The Dissertation Committee for Samuel Vincent Scarpino certifies that this is the approved version of the following dissertation:

Applying Mathematical and Statistical Methods to the Investigation of Complex Biological Questions

Committee:

Mark Kirkpatrick, Supervisor

Lauren Ancel Meyers, Supervisor

Thomas Juenger

Michael Ryan

Sara Sawyer

Applying Mathematical and Statistical Methods to the Investigation of Complex Biological Questions

by

Samuel Vincent Scarpino, B.S. Bio.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas at Austin

August 2013

Dedication

I dedicate this dissertation to my parents Betty and Philip Scarpino.

Acknowledgements

I've taken the elevator to my office on the sixth floor of Patterson Labs 7,236 times. Numbers like these are associated with all the people, places, and things in our lives. Those numbers provide perspective on the role that each has played. It is this perspective that I will try to provide as I acknowledge everyone and everything that has made this Ph.D. a reality.

To my parents, I can't imagine the time, energy, and yes money required to raise a child who earns a Ph.D. All of my successes can be traced to your unwavering support of my endeavors and commitment to my education. For a specific number, I remember drinking at least one quart of milk per day during the last three years of high school. At $2.00 per gallon that works out to around $547, see Dad I told you it wasn't that much. My brother Daniel has visited seven times since I've been in Austin. That's more than anyone else. I'm incredibly fortunate to have him as a brother; his love of adventure and willingness to try anything has been a constant source of joy.

Lynda Delph gave me my first exposure to research science and academia. I decided to pursue a Ph.D. in evolutionary biology because of my work in her lab and the mentorship I received from Lynda, Chris Herlihy, and Mandy Brothers. Lynda's best piece of advice was that choosing the right advisor is the most important aspect of succeeding in graduate school. My Ph.D. advisors Mark Kirkpatrick and Lauren Meyers are testaments to that saying. Not only have they have supported my research, but they have constantly provided me with the opportunities needed to find success after graduate school. Learning from them and counting them as collaborators has been an honor.

In graduate school, your closest colleagues, collaborators, and advice givers are often also your closest friends. For example, my friend, collaborator, and office-mate

Rafael Guerrero and I have sat next to each other for over 10,300 hours during the past six years. He and I have collected fish in Mexico, attended workshops in Switzerland, and hiked to lost cities in Colombia. He's a talented scientist and his advice has been invaluable. I will miss our conversations about science in front of the white board in 653. Dave Des Marais and Robin Hopkins taught me everything I know about bench work. They have been a constant source of support and a pair of wonderful friends. Robin and I traveled together to my first professional conference and since then she has remained an unparalleled source of friendship and advice. With Dave, I've eaten at all of the top ten BBQ places in Texas, climbed the hardest climbs of my life with him holding the other end of the rope (not a metaphor), and enjoyed his constant reminders that we don't work in a "push button" field. Aside from my Ph.D. advisors, Ned Dimitrov has shaped my philosophy of science more than anyone else. It has been a pleasure counting him as a friend and collaborator. Damien Caillaud was my first post-doc mentor as a graduate student. He has shaped my thoughts on science and my perspective on life throughout graduate school. "It's been six years and I've done something."

Although they are small in number, I do have friends that are not collaborators (that's not to say they aren't biologists). However, my one close friend in Austin, who is not a biologist, is Alejandro Puyana. I've played 198 games of pool against Ale and have lost 179. He and I are collaborators, having thrown out some 960 base runners trying desperately for first base. My roommates Liz Milano and Thomas Keller conspired with me on countless challenges, projects, and schemes, all the while providing emotional support through the ups and downs of life. To all my other friends, especially Jesse Lasky, Kevin Bohannon, and Anthony Dee, thank you for providing me with so much joy and happiness these past six years.

Tom Juenger provided a welcoming lab environment and without his support and generosity, the bench portion of my dissertation could not have happened. Simply put, Tom has greatly advanced the caliber and impact of my work. When projects inevitably fail, it's the dissertation committee that picks up the pieces. Aside from Mark, Lauren, and Tom, Mike Ryan and Sara Sawyer have guided me throughout the Ph.D. process. Their advice and support has been fantastic. If I have any regrets in graduate school, it's not having spent more time learning from my committee. I'd like to acknowledge the R programming language, which is behind at least one analysis in very paper I've worked on, and Dan Bolnick for first teaching me the language. Francisco Garcia-De-Leon, a.k.a Paco, spent six weeks with me fishing in Mexico. His friendship, advice on *Xiphophorus*, and generosity in securing collection permits has been critical to the success of our work in that system. I've also been incredibly fortunate to have received advice and guidance from a wonderful set of colleagues and collaborators, especially: Manfred Schartl, Marilyn Felkner, Jeff Taylor, Don Levin, Sally Otto, Patrick Hunt, Roz Eggo, Janine Regneri, Giovanni Petri and all the past/current members of the Kirkpatrick and Meyers labs. The quality of my science and the quality of my life has been greatly advanced by this network of people.

I've been privileged in earning fellowships that afforded me the flexibility to work on a broad range of research questions. The National Science Foundation awarded me a Graduate Research Fellowship, which covered tuition and stipend for half of my dissertation work. As an incoming first-year, I earned a Houston Livestock and Rodeo Fellowship and Dean's Excellence award. The Graduate School at UT Austin awarded me a continuing fellowship for my final year of work. I've also worked as a research assistant for the division of statistics and scientific computation, Lauren Meyers, and Mark Kirkpatrick.

I would also like to acknowledge funding support from the following sources: a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-1110526) to SV Scarpino and M Kirkpatrick, a National Institute of Health Models of Infectious Disease Agent Study grant to LA Meyers and AP Galvani, a National Science Foundation grant (DEB-0819901) to M Kirkpatrick, a National Science Foundation grant (DEB-0546316) to TJ Juenger, a graduate student research grant from the Ecology, Evolution, and Behavior Graduate Program, and a graduate school professional development award. I am grateful to the Mexican federal government and the state of Veracruz for permission to collect *X. maculatus* and *X. hellerii* (permit number DGOPA.00322.25011.-0100). Ron Walter and the Xiphophorus Genetic Stock Center at Texas State University provided useful information and helpful suggestions during the initial phases of my research on *Xiphophorus*.

Finally, I would like to thank Laura Beerits. I started dating Laura exactly one year ago today. Her support and companionship during my final year of school has been a constant source of joy. Finishing your Ph.D. is exciting, but also stressful and tumultuous. I cannot thank her enough for sharing and vastly improving this experience.

Applying Mathematical and Statistical Methods to the Investigation of Complex Biological Questions

Publication No. _____

Samuel Vincent Scarpino, Ph.D.

The University of Texas at Austin, 2013

Supervisors: Mark Kirkpatrick

Lauren Ancel Meyers

The research presented in this dissertation integrates data and theory to examine three important topics in biology. In the first chapter, I investigate genetic variation at two loci involved in a genetic incompatibility in the genus *Xiphophorus*. In this genus, hybrids develop a fatal melanoma due to the interaction of an oncogene and its repressor. Using the genetic variation data from each locus, I fit evolutionary models to test for coevolution between the oncogene and the repressor. The results of this study suggest that the evolutionary trajectory of a microsatellite element in the proximal promoter of the repressor locus is affected by the presence of the oncogene. This study significantly advances our understanding of how loci involved in both a genetic incompatibility and a genetically determined cancer evolve. Chapter two addresses the role polyploidy, or whole genome duplication, has played in generating flowering plant diversity. The question of whether polyploidy events facilitate diversification has received considerable attention among plant and evolutionary biologists. To address this question, I estimated the speciation and genome duplication rates for 60 genera of flowering plants. The

results suggest that diploids, as opposed to polyploids, generate more species diversity. This study represents the broadest comparative analysis to date of the effect of polyploidy on flowering plant diversity. In the final chapter, I develop a computational method for designing disease surveillance networks. The method is a data-driven, geographic optimization of surveillance sites. Networks constructed using this method are predicted to significantly outperform existing networks, in terms of information quality, efficiency, and robustness. This work involved the coordinated efforts of researchers in biology, epidemiology, and operations research with public health decision makers. Together, the results of this dissertation demonstrate the utility of applying quantitative theory and statistical methods to data in order to address complex, biological processes.

Table of Contents

List of Tables

List of Figures

Introduction

Perhaps the greatest achievement in the past two decades has been our ability to gather and store vast amounts of data. For example, since I entered graduate school in 2007, the number of unique sequences on GenBank has doubled from 80 to 160 million and according to IBM, perhaps 90% of all accessible data today was created in the past two years. While there have been successes, this explosion in data has not facilitated a golden age of discovery in biology. This seeming failure of big data has revealed two key insights, (1) data in the absence of testable hypotheses, quantitative theory, and statistical methods is rendered worthless and (2) gathering the right data for the question is more important than the quantity gathered. In this thesis, I develop and apply statistical and theoretical methods to address three important questions in biology. All three chapters involve collecting or aggregating data, constructing quantitative models, and fitting these models to data using powerful statistical methods. The first two chapters pertain to Darwin's mystery of mysteries, speciation. In these chapters, I demonstrate how genes involved in a genetic incompatibility have evolved and the role whole genome duplication has played in generating flowering plant diversity. For chapter 1, I collected molecular and field data on *Xiphophorus* fishes. In the final chapter, I develop a powerful, efficient method for designing disease surveillance networks. The motivation for chapter 3 was to improve the quality of data collected for monitoring infectious diseases.

Chapter 1 presents evidence that two genes involved in a genetic incompatibility in the genus *Xiphophorus*, a group of freshwater fishes, are coevolving. In this genus, inter-specific hybrids often develop a lethal melanoma due to the interaction of two loci, an oncogene and its repressor. Genetic incompatibilities that function as a two-locus,

two-allele trait are referred to as a Bateson-Dobzhansky-Muller (BDM) incompatibility (Bateson 190, Dobzhansky 1936, Muller 1942). Despite the importance of BDM incompatibilities to speciation, *Xiphophorus* remains the only vertebrate system where a BDM incompatibility exists and the identity of both genes is known (Presgraves 2010). I utilized this situation to study the evolutionary forces acting on loci involved in a genetic incompatibility. To investigate the evolutionary relationship of these genes, I cloned and sequenced the repressor locus in 25 *Xiphophorus* species and an out-group. With these data, and presence/absence information on the oncogene taken from the literature, I used phylogenetic methods to test for co-evolution. The results of this study suggest that the evolutionary trajectory of a microsatellite element in the proximal promoter of the repressor locus is affected by the presence of the oncogene. The relevant, genus-level variation at the oncogene is presence/absence, with some species lacking the locus entirely. Three results of this study are of interest to the broader, evolutionary biology community: 1) both structural and regulatory changes can be involved in a genetic incompatibility, 2) microsatellites can play non-neutral roles in the genome, and 3) genetic diseases, in this case melanoma, can be established by positive selection.

Chapter 2 addresses a long-standing question in plant speciation. Namely, what is the relationship between polyploidy, whole genome duplication, and flowering plant diversity? A number of studies have demonstrated a positive relationship between polyploidy and species diversity, leading some to surmise that polyploids may hold an evolutionary advantage over related diploids (Otto and Whitten 2000; Soltis et al. 2003). However, three recent studies that estimated the speciation rate of diploids and polyploids found either evidence for no difference in the speciation rate of diploids and polyploids, Meyers and Levin (2006) and Wood et al. (2009), or for a decrease in the speciation rate of newly formed polyploids, Mayrose et al. (2011). These findings have led to the

proposal that polyploids may, more-often-than-not, be evolutionary dead-ends (Arrigo and Barker 2012). This dichotomy between polyploids as drivers of diversity and polyploids as evolutionary dead-ends, motivated me to investigate the relationship between polyploidy and flowering plant diversity in a broad, comparative context. To address this question, I fit stochastic birth/death models to ploidal level distributions from 60 flowering plant genera. The results suggest a middle ground between the two extreme evolutionary roles proposed for polyploids. First, I show that a simple null model is statistically supported for 55 for the 60 genera included in this study. In this null model, polyploidy is irreversible and con-generic diploids and polyploids are constrained to speciate at equal rates. Second, when the equal speciation rate constraint is relaxed, diploids, as opposed to polyploids, hold a speciation advantage. However, diploids only achieve this evolutionary advantage when counting tetraploid descendants towards their net speciation rate. I also establish that an evolutionary advantage of polyploids is not necessary to account for two common observations, a correlation between the polyploidization rate and the mean genus self-fertilization rate and a marked increase in species richness due to polyploidy. This study represents the broadest comparative analysis to date of the effect of polyploidy on flowering plant diversity.

In chapter 3, I integrate methods in operations research, computer science, and epidemiology to develop an algorithm for constructing disease surveillance networks. Gathering reliable and informative epidemiological data is critical for disease surveillance and public health decision-making during outbreaks. Yet despite its importance, the design of surveillance networks has received surprisingly little attention in the theoretical literature. The method I developed is a data-driven, optimization approach for constructing geographic surveillance networks, which proceeds in three steps. First, historical hospitalization records and existing surveillance reports are used to

simulate thousands of realistic surveillance sites. Second, the desired number of locations are selected for inclusion in the surveillance network using a submodular optimization routine. This step is necessary because an exhaustive search of all possible networks is computationally infeasible. Third, the network is compared to the performance of existing surveillance networks and those constructed using a previously published method (Polgreen et al. 2009). The results demonstrate that networks constructed using this new method are more efficient, robust, and informative than existing surveillance networks and those designed using established methods. Recently, this method was used by the Texas Department of State Health Services to evaluate and augment their influenza surveillance network. Lastly, this work speaks more broadly to the utility of inter-disciplinary research, bringing together researchers in biology and operations research with public health decision makers.

Chapter 1: Evolution of a genetic incompatibility in the genus *Xiphophorus*[1]

Abstract

Genetic incompatibilities are commonly observed between hybridizing species. Although this type of isolating mechanism has received considerable attention, we have few examples describing how genetic incompatibilities evolve. I investigated the evolution of two loci involved in a classic example of a Bateson-Dobzhansky-Muller (BDM) incompatibility in Xiphophorus, a genus of freshwater fishes from northern Central America. Hybrids develop a lethal melanoma due to the interaction of two loci, an oncogene and its repressor. I cloned and sequenced the putative repressor locus in 25 Xiphophorus species and an outgroup species, and determined the status of the oncogene in those species from the literature. Using phylogenetic analyses, I find evidence that a repeat region in the proximal promoter of the repressor is coevolving with the oncogene. The data support a hypothesis that departs from the standard BDM model: it appears the alleles that cause the incompatibilities have coevolved simultaneously within lineages, rather than in allopatric or temporal isolation.

Introduction

What forces drive the evolution of postzygotic isolating mechanisms? Despite the importance of this question to speciation, we in fact know quite little about how

[1] Considerable portions of this chapter were published as Scarpino SV, Hunt PJ, Garcia-De-Leon FJ, Juenger TE, Schartl M, and Kirkpatrick M. *in press* Evolution of a genetic incompatibility in the genus Xiphophorus. Molecular Biology and Evolution doi:10.1093/molbev/mst127. **Contributions** - Conceived and designed the experiments: SVS PJH TEJ MK. Performed the experiments: SVS PJH. Analyzed the data: SVS MK. Contributed reagents/materials/analysis tools: SVS FJGDL MS. Wrote the paper: SVS PJH FJGDL TEJ MS MK.

postzygotic isolation originates (Presgraves 2010). A key idea is the Bateson-Dobzhansky-Muller (BDM) hypothesis, which posits that two loci diverge in populations that are geographically isolated (Bateson 1909; Dobzhansky 1936; Muller 1942). On secondary contact, hybrids are produced that have novel combinations of alleles that reduce fitness. This hypothesis is appealing because it generates an adaptive valley between two populations without requiring either of them to cross the valley during their evolutionary history.

Several genetic incompatibility systems involving multiple interacting loci have now been discovered and interpreted as BDM incompatibilities (Coyne 1992; Presgraves et al. 2003; Presgraves 2010). There are, however, important gaps in our understanding of how these systems have evolved. To date there has been only a single test of the suggestion made by Bateson, Dobzhanksky, and Muller that incompatibilities arise by one substitution occurring in each of two allopatric populations. Contrary to that scenario, Cattani and Presgraves (2009) showed that incompatibilities between *Drosophila mauritiana* and the closely related *D. sechellia* and *D. simulans* arose by two or more substitutions in a single lineage that made it incompatible with the ancestral genotype. A second limitation to our current understanding regards how frequently different forces drive the evolution of incompatibilities. Incompatibilities can evolve as the result of divergent selection (Schluter and Conte 2009), genetic conflict between different parts the genome (Gavrilets 2003; Presgraves 2010), or mutation and random genetic drift (Gavrilets 2004). It is also uncertain whether loci that participate in BDM incompatibilities are typically fixed for the alternative alleles or are polymorphic (Cutter 2012). Finally, there is controversy over whether the genetic changes are most often coding, regulatory, or structural in nature (Hoekstra and Coyne 2007).

The genus *Xiphophorus* provides a remarkable opportunity to investigate how loci involved in a genetic incompatibility have evolved. Crosses between two *Xiphophorus* species provided the first example of a BDM incompatibility (Kosswig 1928; Gordon 1931). *Xiphophorus maculatus* and *X. hellerii* are sympatric, and they occasionally hybridize in nature (Kallman and Kazianis 2006; Meyer et al. 2006; Rosenthal and Garcia-De-Leon 2011; Kang et al. 2013). *Xiphophorus maculatus* individuals carry an oncogene and its repressor locus, both of which are absent or non-functional in *X. hellerii*. In experimental crosses, backcross hybrids segregate for both genes, with about one quarter developing a spontaneous and lethal melanoma. Crossing experiments concluded that this hybrid melanoma is inherited as a two-locus, two-allele trait (Ahuja and Anders 1976). This genetic system, now known as the Gordon-Kosswig cross, has been studied extensively as a laboratory model for melanoma (Meierjohann and Schartl 2006). The same incompatibility system has been documented in crosses between five additional pairs of *Xiphophorus* species (Schartl 2008). *Xiphophorus* remains one of the few systems in which the genetic basis of hybrid incompatibility is known (Coyne 1992; Wu and Ting 2004).

One of the interacting genes in the *Xiphophorus* incompatibility is an oncogene, the *Xiphophorus melanoma receptor kinase* (*xmrk*) gene (Wittbrodt et al. 1989). The *xmrk* gene is unique to *Xiphophorus*, and it arose by a duplication from an epidermal growth factor gene during the diversification of the genus (Weis and Schartl 1998). Critically, not all *Xiphophorus* species have a copy of the *xmrk* locus, and it has apparently been gained and lost multiple times (Schartl 2008). In *X. maculatus* and closely related species that carry *xmrk*, the locus maps to a recombining region of the sex chromosomes (Wittbrodt et al. 1989).

While *xmrk* is known to cause melanoma, it also has effects that may at times be beneficial. Males that carry *xmrk* show increased pigmentation, a phenotype favored by females in mating (Fernandez and Morris 2008). These males are also larger and more aggressive (Fernandez 2010; Fernandez and Bowser 2010). Consequently, sexual selection may favor the presence of *xmrk*, at least in some lineages (Fernandez and Bowser 2010).

The second actor in the incompatibility is a gene whose phenotypic effect is to repress the melanoma. The repressor maps to a region of an autosome (linkage group V) that contains several; genes (Kazianis et al. 1996; Kazianis et al. 1998). Three lines of evidence suggest the repressor is the gene *cdkn2a/b* (Kazianis et al. 1998). First, *cdkn2a/b* is homologous to the mammalian melanoma suppressor locus *CDKN2* (Kazianis et al. 1999). Variation in the promoter and coding regions of this gene are implicated in both the occurrence and progression of melanoma and other cancers (Merlo et al. 1995; Merbs and Sidransky 1999). Second, the two species involved in the Gordon-Kosswig cross, *X. maculatus* and *X. hellerii*, differ in both the coding and promoter regions of *cdkn2a/b* (Kazianis et al. 2000). Third, there is some evidence that *cdkn2a/b* expression in tumor cells in *Xiphophorus* is affected by these differences (Kazianis et al. 1999; Kazianis et al. 2004; Butler et al. 2007). In this chapter I refer to *cdkn2a/b* as the oncogene repressor, but definitive proof of that function using a transgenic construct with a loss-of-function mutant has not yet been obtained.

This situation motivated me to investigate whether *cdkn2a/b* has coevolved with *xmrk*. One plausible scenario, for example, was that *xmrk* might have spread by positive selection (Futreal et al. 2004; Fernandez and Morris 2008), which would then drive the evolution of *cdkn2a/b* to remediate its deleterious effects. To investigate the coevolution hypothesis, I gathered data on the promoter and coding region

of *cdkn2a/b* across the genus *Xiphophorus*. I used these data to test for correlated evolution in the two genes. Specifically, I asked if the gain or loss of *xmrk* is associated with changes in either the coding or regulatory regions of *cdkn2a/b*.

I find evidence for rapid evolution caused by positive selection in the first exon of *cdkn2a/b*, but the pattern of evolution is not correlated with *xmrk* status (presence or absence). Features of the *cdkn2a/b* promoter, however, do correlate with *xmrk* status across the phylogeny. The length of a region in the proximal *cdkn2a/b* promoter is positively associated with the presence of *xmrk*. This region contains several repetitive elements, which I will refer to as the "promoter repeat region" and abbreviate as *PRR*; see Methods for an exact definition of this region. I also find evidence of a correlation between the length of this region and *xmrk* status across three populations of *X. maculatus* that is consistent with the between-species pattern. These results suggest that the oncogene and its repressor have evolved together within lineages. Further, it seems they may coevolve simultaneously, rather than in spatial or temporal isolation as suggested by the BDM hypothesis.

Results

Positive selection on the coding region of cdkn2a/b

The coding region of *cdkn2a/b* has experienced multiple bouts of positive selection. The best-fit model estimated using GABranch shows two selection regimes (Figure 1 and Table 1). That model suggests that on 12 branches of the phylogeny there has been purifying selection ($d_N/d_S = 0$), while on the other 17 branches there has been positive selection ($d_N/d_S = 3.59$). These results give strong support for multiple bouts of amino acid substitution by positive selection.

I then tested whether patterns of evolution in the coding region of *cdkn2a/b* correlate with the presence of *xmrk*. First, I asked if the presence of *xmrk* correlates with the mode of selection (positive or purifying) on terminal branches of the *cdkn2a/b* gene tree. That relationship is not significant (χ^2 = 2.34, p = 0.194 as determined by randomization). I then used stochastic character mapping to perform ancestral state reconstruction across the entire tree ((Bollback 2006), as implemented in the R package Diversitree v. 0.9-6 (FitzJohn 2012)). I did not see an association between positive selection and either the presence of *xmrk* or changes in its status (i.e. gain or loss) (p = 0.99). Last, I looked for a correlation between the rate of evolution of the coding region of *cdkn2a/b* and the presence of *xmrk*. I found no evidence of an association (p = 0.2, TraitRate (Mayrose and Otto 2010)).

Number of regimes	ΔAIC	d_N/d_S	# of branches
1	-4.1	1.18	29
2	**–**	**0**	**12**
		3.59	**17**
3	-9.27	0	12
		0.79	1
		1,000.	16

Table 1 – Results from GABranch.

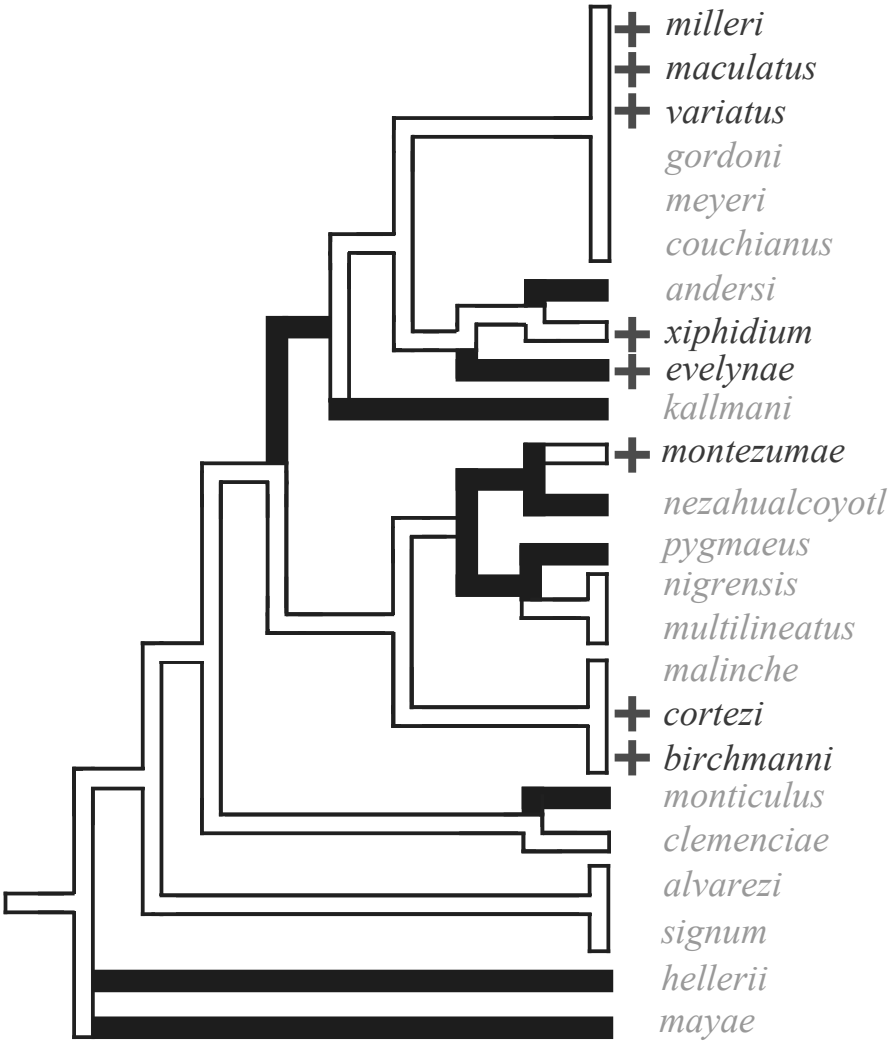

Figure 1 – Positive evolution in the coding region of *cdkn2a/b*.

Shown is the consensus gene tree for the first exon of *cdkn2a/b*. Branches that are colored show statistically significant support for a d_N/d_S ratio that is great than one. The tips are labeled by species, with plus signs and black letters indicating species that have *xmrk* and gray showing species that lack it.

The promoter of cdkn2a/b coevolves with xmrk

I found that the length of the *PRR* of the putative repressor gene *cdkn2a/b* evolves in a correlated way with *xmrk* status. Species with *xmrk* on average have a longer *PRR* than those without. This region was first identified in a comparison between the species in the Gordon-Kosswig cross: *X. maculatus* carries *xmrk* and has a substantially longer promoter than *X. hellerii*, which does not carry *xmrk* (Kazianis et al. 1999; Kazianis et al. 2004). Looking across the entire genus, I find that species with *xmrk* have significantly longer promoters than those without ($p = 0.007$, Wilcoxon Mann-Whitney rank sum test, $z = -2.64$). The results were also significant when *X. maculatus* is removed from the analysis ($p = 0.017$, Wilcoxon Mann-Whitney rank sum test, $z = -2.33$). That analysis does not account for potential phylogenetic dependencies, however, and so I next attempted to control for them.

Figure 2 shows the distribution of the *PRR* length and *xmrk* across the genus. The figure, which is based on just one of the phylogenies used in the following analyses, shows several cases that suggest when closely related species differ in their *xmrk* status, the species that carries *xmrk* tends to have a longer *PRR* at *cdkn2a/b*. Testing for the significance of that pattern is complicated by two issues: uncertainty in the phylogeny and introgression between species. Here I describe how I accounted for phylogenetic uncertainty, and will return to the issue of introgression in the Discussion.

For a given phylogeny, I used BayesTrait Continuous to find the likelihood for a model in which evolution of *xmrk* and the *PRR* are correlated. I averaged those likelihoods over the 1,000 most likely phylogenies (accounting for more than 99% of the credible interval of phylogenetic space), weighting each value by the likelihood of the phylogeny. Using a dataset consisting of six loci to estimate the phylogenies, I reject the null hypothesis that the *PRR* length and the oncogene are evolving independently ($p =$

0.0043, χ^2 = 3.56, d.f. = 1, likelihood ratio test). I also reran the analysis with a recently published dataset that has thirteen loci. The result again shows significant support for correlated evolution (p = 0.026, χ^2 = 2.01, d.f. = 1, likelihood ratio test).

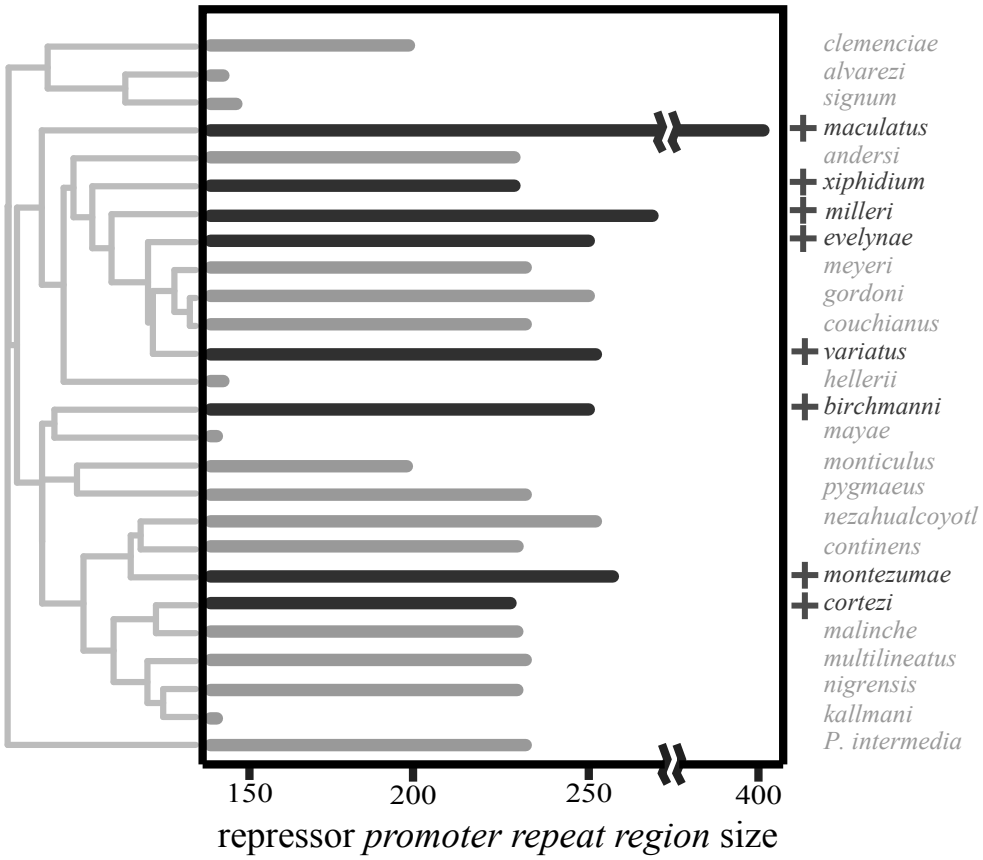

Figure 2 – The phylogenetic distribution of the size of the *PRR* for the putative repressor, *cdkn2a/b*.

At left is a phylogenetic tree that has a high likelihood. The center box shows the *PRR* length. The species names at right contain a plus sign and are black for those that carry *xmrk* and grey for those without *xmrk*.

I then asked if these results might be driven by species with unusual promoters. The *PRR* in *X. maculatus* is much longer than that in any other species. I therefore removed *X. maculatus* from the data set and reran the analyses just described. There is again strong support for correlated evolution ($p = 0.0071$, $\chi^2 = 3.114$, d.f. = 1, likelihood ratio test). I then replaced the actual *PRR* length in *X. maculatus* by the mean promoter length across all species in the genus, and again found significant support for correlated evolution ($p = 0.0048$, $\chi^2 = 3.45$, d.f. = 1, likelihood ratio test).

Finally, I fit the BayesTrait Continuous model to three *Xiphophorus* phylogenies that were recently estimated from genomic-scale data (Cui et al. 2013). Averaging the likelihood for the model across those three phylogenies, I find that support for the correlated evolution model is not quite significant ($p = 0.055$, $\chi^2 = 2.899$, d.f. = 1, likelihood ratio test). Taken together, these results provide support for the hypothesis of correlated evolution of the oncogene (*xmrk*) and the *PRR* of its putative repressor (*cdkn2a/b*).

There are, however, caveats to that conclusion. First (and most obvious), the strength of the conclusion depends in part on which dataset is used to estimate the phylogeny. Second, there is evidence of introgression between species (Meyer et al. 2006; Schumer et al. 2012; Kang et al. 2013; Cui et al. 2013). Consequently, even the true species phylogeny may not accurately reflect the evolutionary history of the two genes I am studying. Last, although there are differences in average promoter lengths, there are also individual species that do not fit the pattern. For example, the platyfish *X. xiphidium* carries the *xmrk* gene but has a shorter *PRR* (244 base pairs) than the non-*xmrk* bearing species *X. gordoni* (251 bp), *X. meyeri* (247 bp), and *X. couchianus* (247 bp). I revisit these issues in the Discussion.

Within-species polymorphism in xmrk and the PRR

The platyfish *X. maculatus* offers another opportunity to test the hypothesis of coevolution of *xmrk* and the *PRR* of *cdkn2a/b*. This species is polymorphic for the presence of *xmrk*, and there is significant variation between populations in its frequency (Figure 3). The coevolution hypothesis lead me to expect a positive correlation across populations in the length of the *cdkn2a/b* promoter and the frequency of *xmrk*.

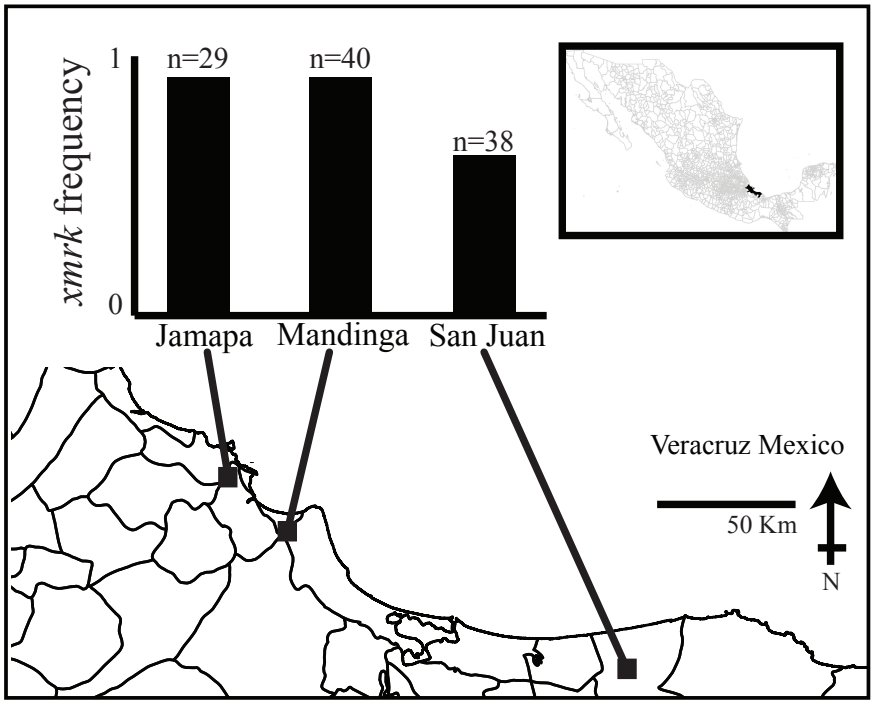

Figure 3 – The frequency of *xmrk* in three *X. maculatus* populations.

The two northern populations, Jamapa and Mandinga, had nearly 100% *xmrk* frequency, while the southern population of San Juan had a significantly lower observed frequency of *xmrk*. The Jamapa and Mandinga populations are likely from the same river drainage, while the San Juan population is from two drainages to the south.

I determined the lengths of the *PRR* in 107 individuals sampled from three populations. In two populations, *xmrk* is near fixation, while in the third population about 60% of individuals have *xmrk* (Figure 3). The population with the lower frequency of *xmrk* also has a significantly shorter mean promoter length ($p < 10^{-7}$, $F = 16.1$, d.f. = 2, ANOVA) (Figure 4).

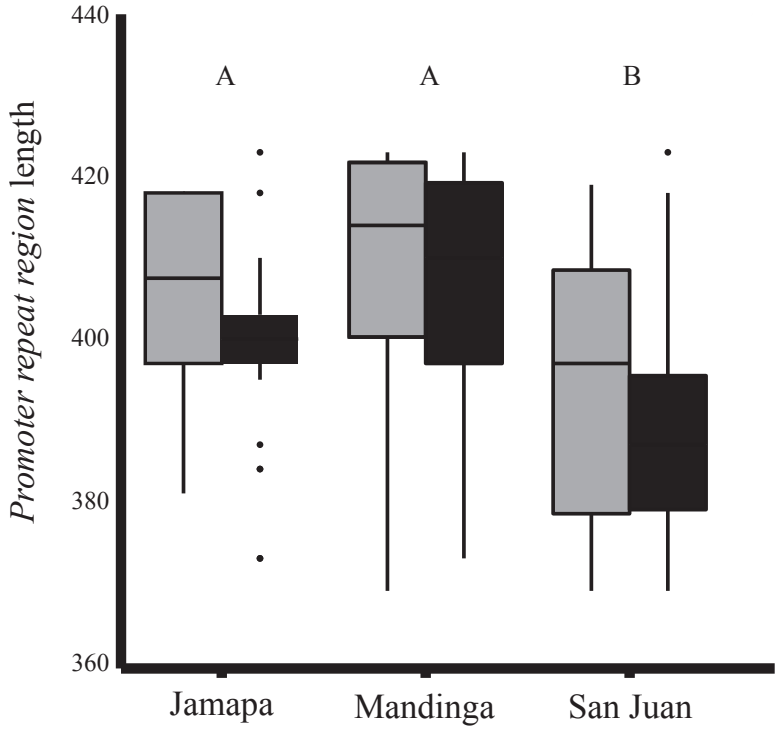

Figure 4 – Average *cdkn2a/b PRR* length correlates with the frequency of *xmrk* across three populations.

Each population is separated into those individuals without *xmrk* (gray) and those with *xmrk* (black). The two populations labeled 'A', Jamapa and Mandinga, had a significantly longer mean *PRR* length than the population labeled 'B', San Juan ($p < 10^{-7}$, $F = 16.1$, df = 2, ANOVA). In the San Juan population, an individual's *xmrk* status is not correlated with the length of its *PRR* ($t = 0.007$, df = 36.0, $p > 0.05$, Welch *t*-test).

In the population with the lower frequency of *xmrk* (San Juan), an individual's *xmrk* status is not correlated with the length of its *PRR* ($p > 0.05$, Welch *t*-test, $t = 0.007$, d.f. = 36.0). That is, there is no evidence that the two loci are in linkage disequilibrium. But because these they are on different chromosomes, however, detectable linkage disequilibrium would not be expected even with strong epistatic selection.

Since I only collected a single population with an intermediate frequency of *xmrk*, I was unable show a significant correlation between *xmrk* frequency and *cdkn2a/b PRR* length across populations. The probability that the pattern I found (the population with lowest *xmrk* frequency has the shortest mean *PRR* lengths) occurred by chance is 1/3. Nevertheless, these data add weight to my earlier conclusions about the coevolution of *xmrk* and *cdkn2a/b*.

Discussion

I have developed three lines of evidence that suggest how *xmrk*, an oncogene, may be coevolving with *cdkn2a/b*, its putative repressor, in the genus *Xiphophorus*. First, I find there have been repeated bouts of amino acid substitution in the first exon of *cdkn2a/b* apparently driven by positive selection. These bouts, however, do not appear to be correlated with *xmrk* status. Second, phylogenetic analyses suggest there has been correlated evolution between the presence of *xmrk* and the length of the *PRR* of *cdkn2a/b*. Third, within one species that is polymorphic for *xmrk*, the *PRR* is longer on average in two populations with high *xmrk* frequencies than it is in a third population with an intermediate frequency of *xmrk*. Taken together, these results support the hypothesis that the *cdkn2a/b* promoter is coevolving with *xmrk*, the oncogene that it is thought to repress.

The evidence from the phylogenetic analyses, however, does have ambiguity. I find significant statistical support for coevolution when accounting for phylogenetic uncertainty by averaging results over the 1,000 most likely phylogenetic trees estimated from two published datasets ($p < 0.03$ for one dataset and $p < 0.005$ for the other). On the other hand, the pattern is only significant at $p < 0.055$ when I use three specific phylogenies based on a much larger genomic dataset (Cui et al. 2013). Even those three phylogenies, however, have uncertainty. For example, they vary slightly depending on the species against which the genome is assembled, and (like all phylogenetic estimates) they are based on an evolutionary model that may not be correct.

An even more complicated issue arises when one considers there is strong evidence of introgression between species (Cui et al. 2013). Consequently, the species phylogeny is likely to differ from the gene trees for *xmrk* and *cdkn2a/b*. Coevolution between the two loci has occurred in the context of their gene trees, not the species phylogenies. Unfortunately, estimates for the gene trees are far less certain than the (already uncertain) species phylogenies.

Given this imprecision, I favor the conclusions that are based on averages across many possible species phylogenies (in the hope that they capture major topological features of the gene trees) rather than relying on any single phylogeny (no matter how well it may be supported). I acknowledge, however, that those conclusions are subject to interpretation.

The oncogene *xmrk* and its repressor have long been cited as a classic example of a two-locus genetic incompatibility (Coyne 1992) (but see Nei and Nozawa (2011) for an alternative view). My results suggest that the *Xiphophorus* incompatibility has evolved in a different way than was originally proposed by Bateson (1909), Dobzhansky (1936), and Muller (1942). They suggested that in one population a

substitution occurs at a first locus, say allele *A* replacing allele *a*, while in an allopatric population an independent substitution occurs at a second locus, say allele *B* replacing allele *b*. Secondary contact and hybridization then produce novel genotypes that carry the two derived alleles (*A* and *B*), and incompatibilities between them cause low fitness. A variant on the BDM hypothesis suggests that the substitution of allele *A* is followed at a later time by substitution in the same population of an allele at a second locus, say allele *C* replacing allele *c*. Following secondary contact, incompatibility then occurs if the derived allele *C* is incompatible with the ancestral allele *a* (Presgraves 2010). This "derived-ancestral" version of the BDM hypothesis is consistent with an analysis of incompatibilities between *Drosophila mauritiana* and the closely related *D. sechellia* and *D. simulans* (Cattani and Presgraves 2009). Central to both versions of the BDM hypothesis is that low-fitness genotypes are never produced in the evolutionary history of either population before they have secondary contact.

My phylogenetic analyses of coevolution of *xmrk* and *cdkn2a/b* suggest that both partners in the incompatibilities have evolved within single lineages, as envisioned in the derived-ancestral version of the BDM hypothesis. But in a departure from that hypothesis, the data further suggest that incompatible genotypes have been produced within single populations as the result of the *simultaneous* (rather than sequential) evolution of the two loci. I find that populations of *X. maculatus* are polymorphic at both loci, suggesting that the two genes coevolve and generate incompatible genotypes even in the absence of secondary contact.

What could cause *xmrk* to spread despite its deleterious effects? Fernandez and Morris (2008) and Fernandez and Bowser (2010) make the fascinating suggestion that *xmrk* has been favored by sexual selection that was sufficiently strong to offset the oncogene's negative effects on viability. A second possibility is that *xmrk*

spread by hitchhiking with the closely-linked macromelanophore locus, which has a function in kin recognition (Franck et al. 2001). A third hypothesis, also consistent with my phylogenetic analyses, is that evolution of *cdkn2a/b* drives evolution of the system. Its promoter might evolve by selection on a pleiotropic effect, by mutation pressure from the microsatellite motif, or simply by drift. Once an allele that acts as a melanoma repressor reaches an appreciable frequency, it enables *xmrk* to invade.

Under all three of these hypotheses, the fitness "valley" between the species caused by the incompatibilities in fact did not exist, at least when they first evolved. For example, under the Fernandez et al. hypothesis, the viability cost of the melanoma was offset by the reproductive advantage that *xmrk* conferred. It is certainly possible that the incompatibilities create a true fitness valley at present, for example because of changed pressures of sexual selection or the ecological environment. Further, all of these scenarios are made more plausible by the fact that the melanoma repressor has a dominant gene action. (Recall that melanomas do not appear until the backcross and later generations of the *maculatus* x *hellerii* cross.) That substantially decreases the negative selection against the incompatibility compared to cases where F_1 hybrids suffer an immediate fitness cost.

How could *xmrk* be gained and lost multiple times across the genus? Even if several independent losses by deletion seem plausible, parallel gains by duplication seem unlikely. My phylogenetic analyses do not rule out the hypothesis that *xmrk* had only a single origin and was lost repeatedly. Another possibility is that multiple gains of *xmrk* have occurred by hybridization and introgression between species. Both phenomena are common in *Xiphophorus* (Meyer et al. 2006; Schumer et al. 2012; Kang et al. 2013; Cui et al. 2013). Introgression could also explain some of the deviations from the positive correlation between *xmrk* and the *PRR* length of *cdkn2a/b*: for example,

some of *Xiphophorus* species that depart from expected *PRR* lengths are known to hybridize in nature (Schartl 2008; Rosenthal and Garcia-De-Leon 2011; Cui et al. 2013).

In this study, I considered *xmrk* to be a binary trait. Detailed molecular analysis, however, has revealed there is additional variation associated with this locus (Weis and Schartl 1998; Schartl 2008; Regneri and Schartl 2011). In laboratory strains of *X. maculatus*, insertion/deletion polymorphism in the promoter and coding region of *xmrk* are associated with variation in melanoma prevalence, progression, and severity (Regneri and Schartl 2011). Thus variation at this locus is more complex than just its presence or absence. My phylogenetic analyses also ignore within-species polymorphisms for the presence of *xmrk*. In fact, polymorphism is present in *X. maculatus* (Schartl 1990), *X. montezumae*, and all other *xmrk* carrying species studied to date (Schartl M, unpublished data). The analyses presented here neglect this polymorphism, and its evolutionary significance (if any) is unknown. However, the phylogenetic analyses are conservative with respect to polymorphism because any decrease in *xmrk* frequency will decrease its effect on the evolution of *cdkn2a/b*. No species scored as *xmrk* negative in my analyses (see Figure 1) has been found to carry a functional copy of the gene.

An important gap in this story is that *cdkn2a/b* has not been validated as the melanoma repressor using expression assays or transgenic constructs. There is, however, support for its role: in the Gordon-Kosswig cross, there are differences in *cdkn2a/b* expression between healthy individuals and those with *xmrk*-induced melanoma (Kazianis et al. 1999; Kazianis et al. 2004; Butler et al. 2007). Further, the cancer phenotype depends on whether an individual carries the *X. maculatus* or *X. hellerii* allele at *cdkn2a/b* (Butler et al. 2007). That observation, however, does not show whether it is

the *PRR* or some other linked region that is responsible. Functional validation of the effects of the *PRR* is an important goal of future research.

How might variation at *cdkn2a/b* affect the melanoma? The length of the repressor's *PRR* could directly modulate expression of the *cdkn2a/b* protein, which then affects expression of *xmrk*. Alternatively, *cdkn2a/b* expression could act as a regulator (either upstream or downstream) at some other point in the melanoma pathway. This second scenario includes the case where *cdkn2a/b* affects expression of other closely linked genes that are inside the mapping region identified for the repressor (Schartl et al. 2013). The *PRR* of *cdkn2a/b* contains a series of predicted transcription factor binding sites. An important question is whether these length differences could in fact alter expression of the melanoma phenotype. That hypothesis is made plausible for one of the elements in the *PRR*, a GT microsatellite, by evidence from another complex disease, asthma. In humans, length variation in a GT repeat region of the *STAT6* promoter is associated with both progression and symptom severity (Gao et al. 2004). The addition of three bases in the GT repeat region is sufficient to explain symptom variation (Gao et al. 2004).

My findings contribute further understanding to three general issues regarding the genetics of postzygotic isolation. Several evolutionary forces have been implicated in the evolution of postzygotic incompatibilities, including positive selection, neutral processes (drift and mutation), and genomic conflict (Presgraves 2010). While the evidence is not definitive, the most plausible interpretation of the *Xiphophorus* system is that this incompatibility evolved by positive selection. Second, I find variation within species for the genetic incompatibility in the form of presence/absence polymorphisms for *xmrk* status and variable lengths of the *cdkn2a/b PRR*. Thus *Xiphophorus* provides another case of a genetically variable postzygotic incompatibility (Cutter 2012). Third,

there is an ongoing discussion of whether postzygotic isolation typically results from evolution in coding, regulatory, or structural features of the genome (Hoekstra and Coyne 2007). In the case of *Xiphophorus*, evolution of the *cdkn2a/b PRR* is likely both structural and regulatory, while presence/absence variation at *xmrk* is an example of structural variation.

I have focused this chapter on the role that *xmrk* and its repressor may play in postzygotic isolation. The system offers rich opportunities to study other key problems in evolutionary genetics. These include the roles played by microsatellites in adaptation, by sexually antagonistic selection in the evolution of sex chromosomes, and by positive selection in the evolution of oncogenes.

The Models, Methods, and Data

Data Collection

I obtained genomic DNA from all species in the genus *Xiphophorus* (except *X. mixei*) and from a closely related outgroup species, *Priapella intermedia*. Tissue was taken from wild-caught individuals and from laboratory stocks that had been established from wild-caught fish. *Xiphophorus malinche* genomic DNA was generously provided by G. Rosenthal. I studied patterns of variation within *X. maculatus* using 107 individuals that I collected from three populations in Veracruz, Mexico in 2008, 2009, and 2011. These populations are: *jamapa* – Rio Jamapa (19° 0'49.90"N 96°14'51.76"W), *mandinga* – Rio Jamapa (19° 0'46.97"N 96° 5'45.45"W), and *san juan* – Rio San Juan (18°20'3.65"N 95°27'36.95"W).

Data on *xmrk* status for all species was taken from Weis and Schartl (1998) and Schartl (2008), who used a variety of Southern blot and probe-based methods

to detect *xmrk*. All species where *xmrk* is present are polymorphic for *xmrk*, with variable frequencies of the locus between populations (Schartl 1990 and Schartl M, unpublished data). I coded all species where *xmrk* has been found at least once as having *xmrk*, which was conservative with respect to the phylogenetic analyses (see the Discussion).

I amplified, cloned, and sequenced the first 1000 bases of coding region of *cdkn2a/b* for all species. Primers spanning the first exon were: For- ACG CCT GGT TCG GTT TTC CT and Rev-GCC TTA TTC ACG GTT CTC AAT C. PCR conditions were: initial denaturing time of 5min at 94°C, 40 cycles of 94°C 30sec, 59°C 30sec, 72°C 30sec and a final elongation step of 5min at 72°C. Cloning followed standard procedures for TOPO TA cloning with pCR 2.1-TOPO and TOP10 chemically competent cells, Invitrogen K4500-01. Sequences were deposited in GenBank under accession numbers KF002384 – KF002407.

The promoters of *cdkn2a/b* in *X. maculatus* and *X. hellerii*, which are the species in the famous Gordon-Kosswig cross that first identified this two-locus incompatibility, have several differences (Kazianis et al. 1999). I focused on the most conspicuous features of this region: a length polymorphism containing a number of repetitive elements. The most striking is a GT microsatellite (Kazianis et al. 1999; Kazianis et al. 2004) that lies approximately 450 bases upstream from the start codon. This GT microsatellite is about 25 bases long in *X. hellerii* and about 170 bases long in *X. maculatus*. Because this GT repeat feature is a microsatellite, it may experience elevated mutation rates.

Pilot sequencing identified a highly variable repeat region in the proximal *cdkn2a/b* promoter that contains the GT microsatellite. (These sequences are deposited in GenBank under accession numbers KF002357 – KF002383.) I refer to this section of

the proximal promoter as the *PRR*. The boundaries of the region are defined by conserved non-repetitive motifs that are given by the primer pair: For- ACA CTA AAT AGC CCT CTA CCA, Rev- CAT AAA CAC CAG ACT GAA ACA C. To obtain precise estimates of its length, I used a fragment analysis. I amplified the *PRR* using the PCR conditions described above but with an annealing temperature of 51.5°C and using fluorescently labeled primers corresponding to the two sequences just stated. Fragment analysis was performed with standard protocols from the Institute for Molecular and Cell Biology's core facility at the University of Texas at Austin.

Models and Analyses

I analyzed the pattern of selection acting on the coding region of *cdkn2a/b* using GABranch (Pond and Frost 2004). I assumed the best-fit model of nucleotide evolution for the data, which was HKY85 (a model with two substitution rates). The method proceeds in three steps. First, I assume that along each branch the rates of nucleotide substitution are chosen from one of B sets of the two substitution rates. With $B = 1$, for example, a single set of two rates pertains to the entire tree, while with $B = 2$ there are two types of branches each with its own pair of values for the two rates. Second, maximum likelihood is used to estimate B (the number of substitution rate sets), the length of each branch on the gene tree, and which of the B sets of substitution rates pertain to each branch. Third, the fit of the model with different values of B are compared, and the optimal value of B is determined using the Akaike Information Criterion. (GABranch uses a version that includes a correction for small sample size, "AICc" (Akaike 1974; Sugiura 1978).) This analysis provides two kinds of information. First, it gives an estimate of d_N/d_S, the ratio of the synonymous to nonsynonymous rates

of substitution, for the first exon of *cdkn2a/b* along each branch. Second, it determines whether this ratio varies across the gene tree.

To determine whether *xmrk* status was associated with bouts of positive selection, I first estimated ancestral states for branches and nodes and inferred the timing of gains and loses using stochastic character mapping. Stochastic character mapping was performed using the Diversitree package (v. 0.9-3) in R (v. 2.15.1) (FitzJohn 2012).

Next, I tested for correlated evolution between the *xmrk* status (its presence or absence) and the length of the *cdkn2a/b PRR*. The evolutionary hypothesis I tested has two components. The first is that *xmrk* has been gained and lost multiple times across the phylogeny. A special case of this model includes the scenario where *xmrk* had a single origin early in the evolution of the genus and was lost multiple times. The second part of the hypothesis is that the *cdkn2a/b PRR* is a continuous trait under stabilizing selection with an optimal length that differs depending on whether *xmrk* is present or absent. Within each of these two regimes (*xmrk* present vs. absent), changing selection pressures and random genetic drift might cause the promoter length to vary in time around the optimum.

The statistical question therefore is whether the *PRR* lengths differ significantly when *xmrk* is present and when it is absent. I tested for that difference using BayesTrait Continuous (Pagel 1994). This model assumes that *xmrk*, a binary trait, evolves as Markov chain and that the *PRR* length, a continuous trait, evolves by Brownian motion with a constant variance. It calculates the likelihood of a correlation between average *PRR* length and *xmrk* status given a phylogeny. To control for phylogenetic non-independence, the method uses a random effects model. I further evaluated the utility of BayesTrait Continuous using simulations of both the correlated and uncorrelated

evolutionary model. The results demonstrate support for the ability of BayesTrait Continuous to favor the correct model of evolution used in the simulation.

I used three approaches for the phylogenetic analyses. First, I estimated phylogenies using sequences from five nuclear loci and one mitochondrial locus included in Meyer *et al*. (1994 and 2006). Trees were sampled and their probabilities calculated using Mr Bayes (Ronquist et al. 2012). I obtained an overall likelihood by weighting each model likelihood by the probability for that tree (Huelsenbeck et al. 2000). The calculations in BayesTrait continuous are computationally fast, which allows me to calculate the likelihood for both the correlated and uncorrelated model on each of the 1000 most likely trees. Second, I repeated the analysis using sequences from eleven nuclear and two mitochondrial loci included in Kang et al. (2013). All of the loci contained in (Meyer et al. 1994; Meyer et al. 2006), except one nuclear locus, were contained in the Kang et al. (2013) dataset. Finally, I used three recently published phylogenies constructed using genomic data (Cui et al. 2013). Data for these phylogenies were obtained from Dryad (Cui et al. 2013). Again, I fit both models, correlated and uncorrelated, to these trees.

Chapter 2: The effect of polyploidy on flowering plant abundance[2]

Abstract

Polyploidy, or whole genome duplication, has been an important feature of eukaryotic evolution. This is especially true in flowering plants, where between 50 – 100% of extant angiosperms have descended from polyploid species. Here, I present a broad comparative analysis of the effect of polyploidy on flowering plant diversity. I examine the widely held hypothesis that polyploid flowering plants generate more diversity than their diploid counterparts, by fitting stochastic birth/death models to observed ploidal frequency data from 60 extant angiosperm genera. My results suggest the opposite, that diploids speciate at higher rates than polyploids, through a combination of simple diploid speciation and tetraploidy. My model is able to account for two common empirical observations without assuming that polyploids have a speciation advantage over diploids: 1) a correlation between polyploidy and species richness and 2) a positive relationship between the polyploid formation rate and the observed self-fertilization rate.

Introduction

Recent genome level data has confirmed the ubiquity of polyploidy in plants (Soltis et al. 2009). Current estimates suggest between 50 – 100% of all extant angiosperms have a polyploid ancestor and that 20 – 50% are recently formed polyploids

[2] Considerable portions of this chapter are in review for publication as Scarpino SV, Levin DA, and Meyers LA. Polyploidy not polyploids shapes flowering plant diversity. **Contributions** - Conceived and designed the experiments: SVS DAL LAM. Performed the experiments: SVS. Analyzed the data: SVS. Contributed reagents/materials/analysis tools: SVS DAL LAM. Wrote the paper: SVS DAL LAM.

(Levin 2002; Soltis et al. 2009). Although little doubt remains about the pervasiveness of polyploidy in flowering plants, there is considerable debate over whether diploids and polyploids differ in speciation and diversification rates.

Otto and Whitten (2000) reported a positive correlation between polyploidy and species richness. This, combined with evidence for polyploid-biased lineage survivorship through the Cretaceous-Tertiary boundary (Fawcett et al. 2009 & Soltis and Burleigh 2009) and anecdotal support for polyploid superiority, has stimulated an ongoing discussion focused on possible biological and genetic (intrinsic) advantages of polyploids relative to diploids, as reviewed by Soltis et al. (2003). One hypothesis is that chromosome doubling itself may alter the ecological tolerances of populations, thereby allowing them to invade new habitats and rapidly establish (Stebbins 1980 and 1985; Levin 1983 and 2002). Another hypothesis is that phenotypic variation arising during polyploidy events enables the invasion of new habitats, mediated by a range of possible genetic and epigenetic mechanisms, as reviewed by Beest et al. (2012).

In contrast, several recent studies estimating speciation and extinction rates of polyploids and diploids have failed to find evidence for a polyploid speciation advantage: Wood *et al*. (2009) found no evidence for higher polyploid diversification rates in twelve angiosperm genera, and Mayrose *et al*. (2011) found that net speciation rates in recently formed polyploids were lower than congeneric diploids. These studies have led some to conclude that polyploids lead to evolutionary dead ends (Arrigo and Barker 2012).

In response to the various hypotheses about polyploidy advantages, Meyers and Levin (2006) proposed that the high frequency of polyploids in flowering plants might simply be an inevitable consequence of the directionality of polyploidy. Ploidal increases are largely irreversible over short evolutionary timescales, and therefore the abundance of polyploids should increase over time in a ratchet-like manner (Stebbins 1971). In their

analysis, Meyers and Levin (2006) fit a simple, deterministic evolutionary model of polyploid evolution to data from ten angiosperm genera. Although this model assumed that speciation in congeneric diploids and polyploids occurred at the same rate, it was still able to produce distributions of ploidal levels that were statistically similar to those observed in nine of the ten focal genera. This suggests that the irreversibility of polyploidy itself may explain the ubiquity of polyploids. These results also imply that the ratchet model should serve as a parsimonious baseline (null model) when considering other possible explanations for polyploid abundance.

Here, I address the three conflicting hypotheses about the evolutionary potential of polyploids—that they are drivers of diversification, evolutionary dead-ends, or neither — by examining the relationship between polyploidy and flowering plant diversity in a broad, comparative context. Specifically, I extend the model introduced in Meyers and Levin (2006) and apply it to data from 60 angiosperm genera to assess whether: (1) the simple, ratchet model can explain ploidal level distributions across this phylogenetically broader set of taxa, (2) there is statistical evidence for differences in the net speciation rates of polyploids and diploids, and, if so, in which direction, and (3) allopolyploidy has contributed more to the formation of new polyploid lineages than autopolyploidy.

In short, I find that the answer to each of these three questions is yes. The simple, ratchet model is supported in 55 of the 60 genera considered, diploids and polyploids appear to speciate at different rates, and allopolyploidy is estimated to occur at nearly twice the rate of autopolyploidy. However, my comparison of diploids and polyploids yielded complicated and surprising results. In terms of net speciation—the evolution of descendant species through a combination of diversification and polyploidization—diploids generate diversity at higher rates than polyploids, largely through the frequent formation of tetraploids. If I exclude polyploidization and consider only speciation

events that maintain parental ploidal level, then diploids and polyploids are statistically indistinguishable. This suggests a nuanced view of the contribution of polyploids to flowering plant diversity, in which polyploids as a whole are neither superior nor inferior to diploids, and the only special evolutionary engine is the formation of tetraploids from diploids—construed perhaps as a diploid speciation advantage or a tetraploid establishment advantage.

My model is also able to account for two empirical observations about polyploids. The first, as described earlier, is a positive correlation between species richness and a genus' average ploidal level. Second, is a positive correlation between degree of self-fertilization and a genus' average ploidal level. Several studies have observed a positive correlation between the self-fertilization rate of a genus and either its rate of polyploid formation (Stebbins 1950 and Soltis et al. 2003) or the extent of polyploidy in the genus (Barringer 2007). Polyploidy rates and self-fertilization may be correlated for a number of biological reasons including minority cytotype avoidance (Levin 1975), a breakdown in self incompatibly systems (Mable 2004), and decreased inbreeding depression in polyploids (Lande and Schemske 1985). I demonstrate that the ratchet model is able to account for both observations, even in the absence of an evolutionary advantage of polyploids.

Results

I first assess whether the Simple Ratchet model—which forces speciation rates to be equal for diploids and higher ploids—can explain the distributions of ploidal levels observed for the 60 genera considered, then use the Complex Ratchet model—which relaxes this assumption—to estimate the rates of diversification, polyploidization, and the

relative contributions of allopolyploidy and autopolyploidy. Finally, I demonstrate that two key empirical observations that have fueled speculation about the evolutionary advantages of polyploids—the positive correlation between polyploidy rates and the mean genus self-fertilization rate, and a marked increase in species richness stemming from polyploidy—are expected to occur even if polyploids are at an evolutionary disadvantage.

Estimating and comparing evolutionary rates

Using the goodness-of-fit test, I find that the Simple Ratchet model of polyploidy evolution introduced in Meyers and Levin (2006) is sufficient to account for the distribution of ploidal levels in 55 of the 60 genera considered (Figure 5), just slightly lower than the 95% of model rejections expected at $\alpha = 0.05$. This model assumes that congeneric diploids and polyploids speciate at equal rates. The five genera for which the model was rejected are *Acacia*, *Geranium*, *Rubus*, *Elymus*, and *Silene*.

Using the Complex Ratchet model, which allows different diversification rates for diploids and higher ploidal species, I estimate and compare these rates. I find statistical support for a diploid speciation advantage (Figure 6). Using phylogenetic group as a random effect, I calculated a phylogenetically-independent difference between diploid and polyploidy net speciation rates: $Diff = (r_d + h_d) - (r_p + h_p)$. The estimated difference was 0.467 (95% CI 0.383 – 0.551, $p<1x10^{-4}$), indicating a diploid advantage. These results were further supported using a Welch's two-sample t-test on paired data (ignoring phylogeny) ($t=13.57$, $p<1x10^{-16}$).

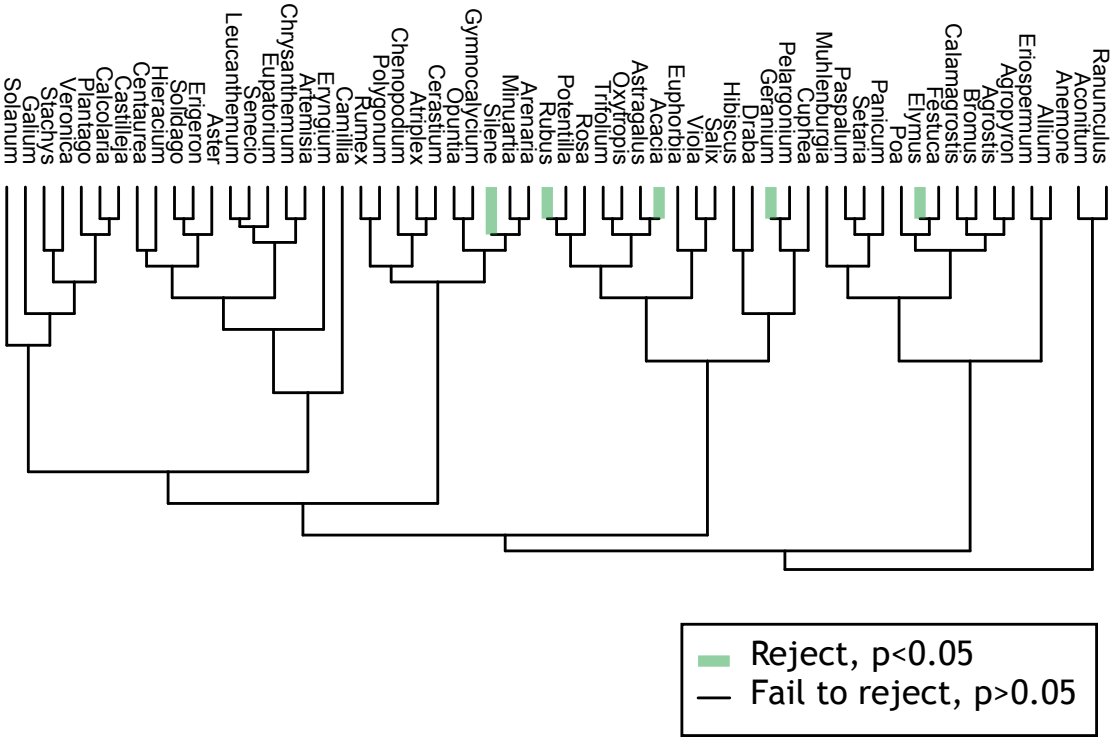

Figure 5 – Support for the Simple Ratchet model.

The phylogeny of genera included in this study, with genera failing to support the Simple Ratchet model indicated with thick green branches. Using a simulation-based, goodness-of-fit test, the Simple Ratchet model was sufficient in 55 out of 60 genera at the $\alpha = 0.05$ level. The five exceptions are *Acacia*, *Geranium*, *Rubus*, *Elymus*, and *Silene*.

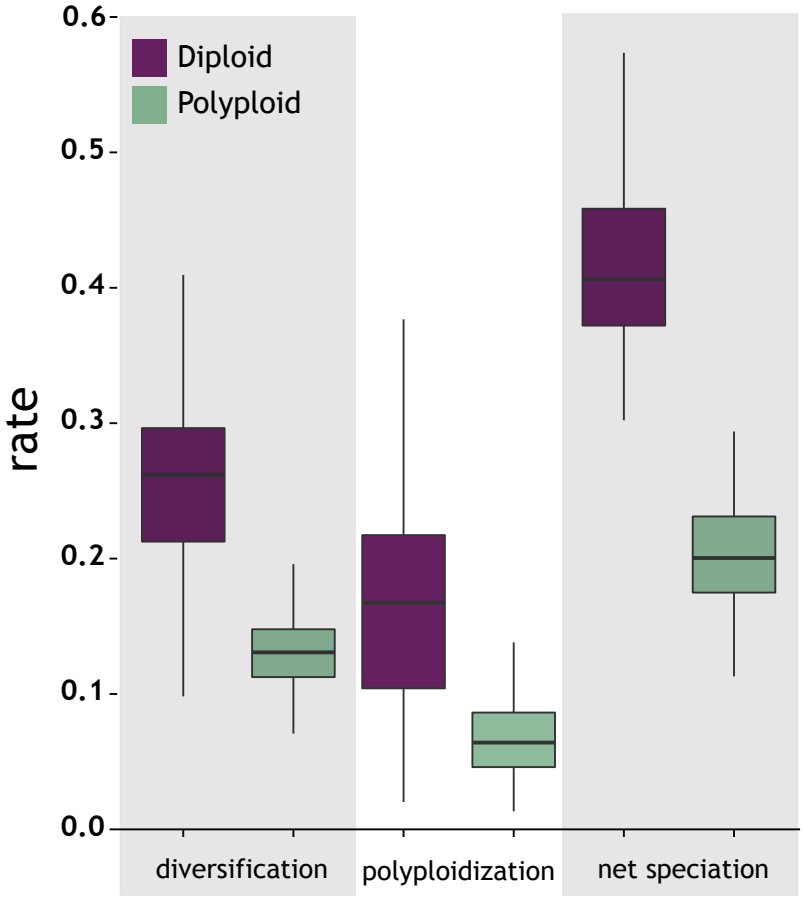

Figure 6 – Diploid speciation advantage.

Boxplots with the median, interquartile, and range for the estimated diversification (r), polyploidization (h), and net speciation rates ($\lambda = r + h - \mu$) from each of the 60 genera. Using phylogenetic group as a random effect, I calculated a phylogenetically independent difference between the mean diploid and polyploid net speciation rates across all 60 genera, $Diff = (r_d + h_d) - (r_p + h_p)$. The estimated difference was 0.467 (95% CI 0.383 – 0.551, $p<1x10^{-4}$). A positive number indicate a diploids advantage. This net speciation rate advantage of diploids stems primarily from the formation tetraploids (diploid polyploidization).

I also estimate that allopolyploidy (as opposed to autopolyploidy) was responsible for the majority of polyploidy events in most genera across (Figure 7). However, this approach suffered from low statistical power to estimate the fraction of autopolyploidy. Most of the statistical information on the fraction of autopolyploids (a) derives from the number of 6n – 14n species. The model assumes diploid foundry and therefore under a model without allopolyploidy, $a = 1$, no species with 6n – 14n genomes can be formed. The vast majority of species considered have ploidal levels less than 6n.

To investigate the effect of extinction, I compared versions of the Complex Ratchet model with and without extinction ($\mu = 0.1$ and $\mu = 0$). Intuitively, extinction caused an increase in the estimates across all speciation rates (Figure 8). This yielded a larger discrepancy between the estimated net speciation rate of diploids and polyploids, driven primarily by differences in polyploidization rates rather than diversification rates, as determined using the proportional increase in diversification rate (Welch t-test comparing $\frac{r_d^{\mu=0.1}}{r_d^{\mu=0}}$ and $\frac{r_p^{\mu=0.1}}{r_p^{\mu=0}}$, $t = 1.325, p = 0.19$) and polyploidization rate: (Welch t-test comparing $\frac{h_d^{\mu=0.1}}{h_d^{\mu=0}}$ and $\frac{h_p^{\mu=0.1}}{h_p^{\mu=0}}$, $t = 5.336, p = 1\text{e-}6$).

I also considered models where the extinction rate was either a free parameter or a function of the diversification rate, r. However, this introduced a substantial amount of uncertainty in all parameter estimates and I was unable to estimate the relative speciation rates of diploids and polyploids.

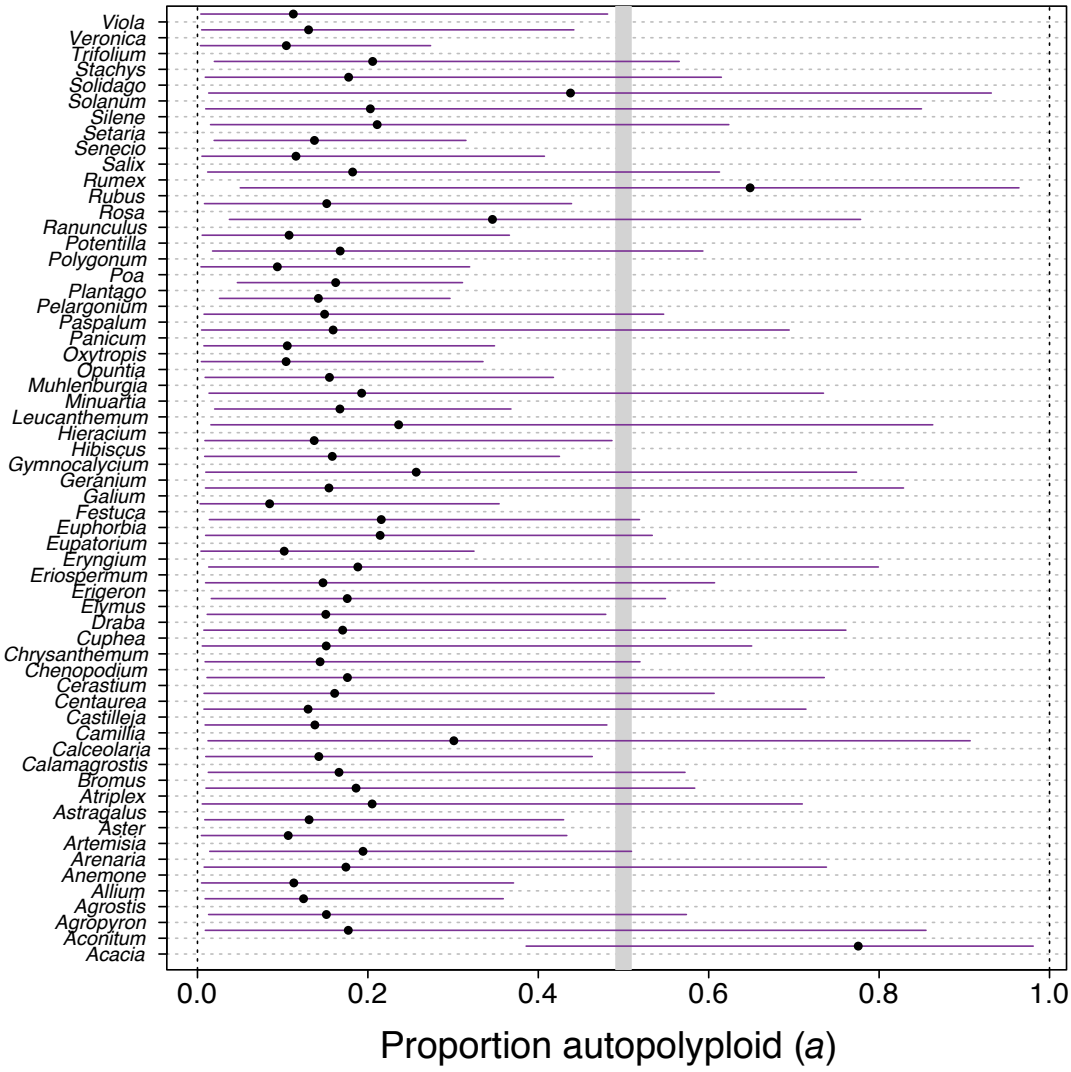

Figure 7 – Proportion autopolyploid.

The estimated proportion of polyploid lineages formed due to autopolyploidy (rather than allopolyploidy) for each genus (a). Error bars represent the 95% Credible Intervals of the posterior distribution. Only *Acacia* and *Rubus* are estimated to have (a) greater than 0.5 (gray vertical bar), indicating that more than half of the polyploidy events are due to autopolyploidy.

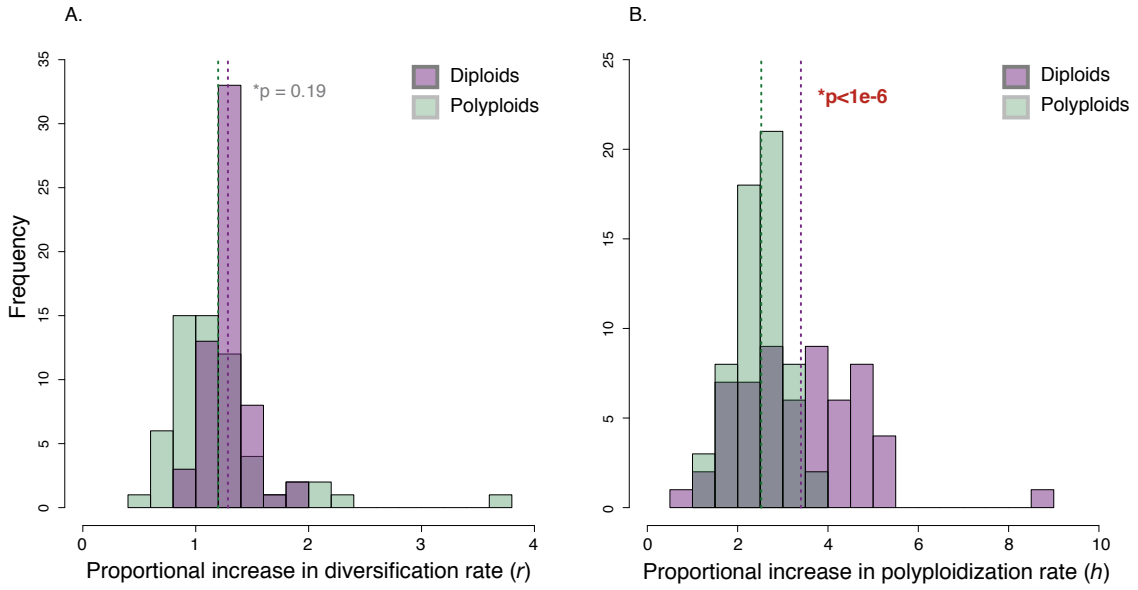

Figure 8 – Extinction's effect on diversification & polyploidization rates.

A proportional increase greater than one indicates that the estimate increases when extinction is explicitly modeled. The diversification rate (r) estimates increase similarly for diploids and higher ploids, while the polyploidization rate (h) estimates increase more for diploids than higher ploids; Welch t-test on the proportional increases in diversification rate: ($t = 1.325, p = 0.19$) and polyploidization rate: ($t = 5.336, p = 1e\text{-}6$).

Polyploidy and self-fertilization rates

Using a previously published data set on the self-fertilization rates of species in naturally occurring populations (Barringer 2007), I compared my evolutionary rate estimates to mean genus self-fertilization rates (Figure 9). I found a positive correlation between the mean self-fertilization and tetraploids formation rates, h_d, ($\beta = 0.432, R^2 = 0.235, p = 0.032$), but no significant relationship between self-fertilization and the polyploid polyploidization rates, h_p.

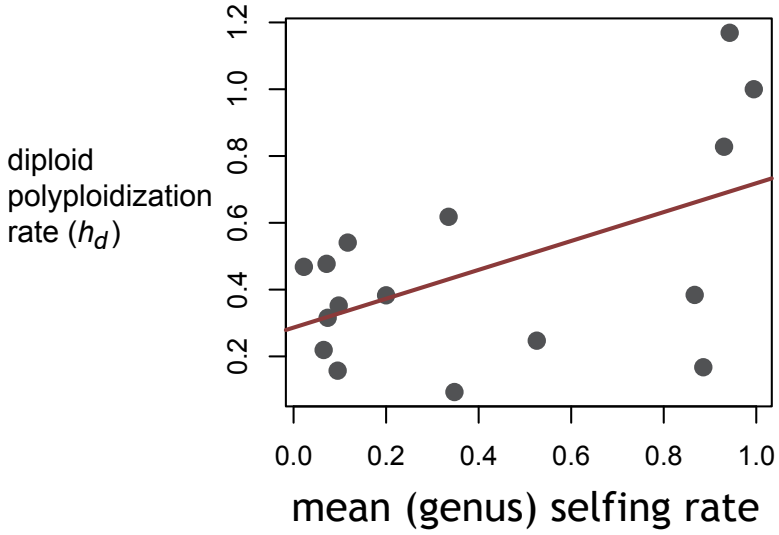

Figure 9 – Polyploidy and self-fertilization.

The relationship between estimated polyploidization rates and mean self-fertilization rate for 16 genera contained in both this study and Barringer (2009). There is a significant, positive relationship between the mean self-fertilization rate and the diploid polyploidization rate (h_d); (standard least-squares regression - $\beta = 0.432$, $R^2 = 0.235$, $p = 0.032$).

Polyploidy and Species Richness

Polyploidization increases diversity by directly generating new species and creating lineages able to undergo further diversification, and thus can fuel increases in species richness even if polyploids diversify at the same rate as (or even slower than) diploids. For example, in *Eriospermum*, the polyploid net speciation rate is estimated to be equal to the diploid net speciation rate, and the Simple Ratchet model projects that polyploid species will grow to dominate the genus (Figure 10). In this way,

polyploidization itself can explain the increased species richness associated with clades containing more polyploids, without invoking an evolutionary advantage of polyploids.

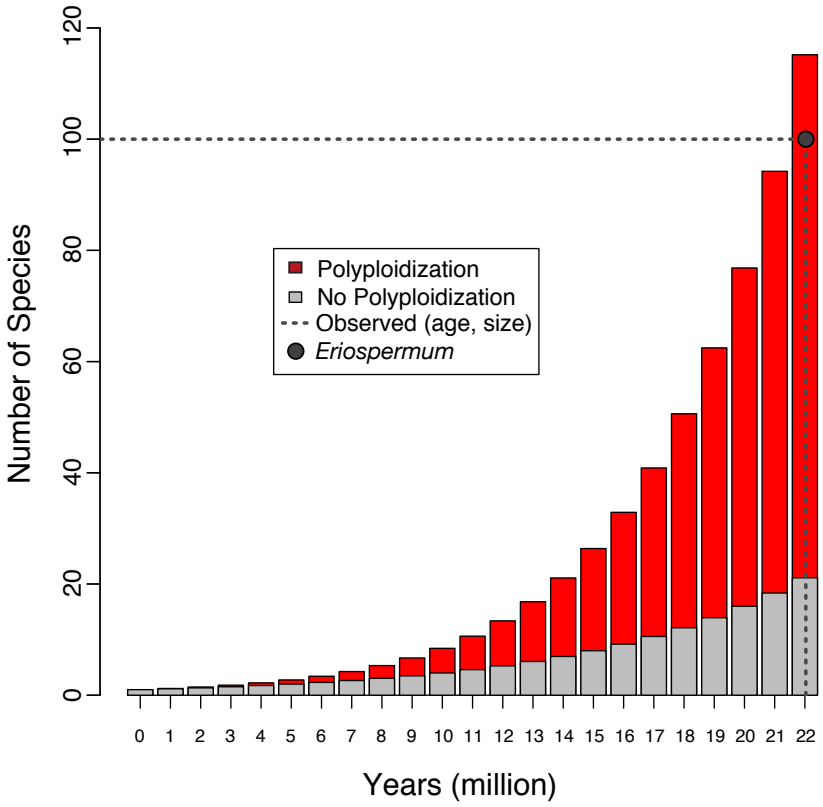

Figure 10 – The correlated ascent of polyploids and total species.

Using the estimated values for the diploid diversification, polyploid diversification and polyploidization rates for *Eriospermum*, I projected the number of species over time (in millions of years), both with (red/dark) and without (gray/light) polyploids. The greater than two-fold difference in the number of species at 22 MY suggests that the Simple Ratchet model can generate the observed correlation between species richness and average ploidal level (without assuming any polyploidy superiority).

Discussion

I have analyzed ploidal distribution data from 60 flowering plant genera representing 25 families across 17 orders and found evidence that diploids diversify faster than their congeneric polyploids. This diploid advantage stems primarily from a higher rate of polyploidization in diploids than polyploids. My model also accounts for two empirical correlates of polyploidy, without assuming a polyploidy speciation advantage: high levels of species richness and high rates of self-fertilization.

In their original paper, Meyers and Levin (2006) introduced a simple quantitative model of angiosperm evolution in which polyploids held no evolutionary advantage over diploids and polyploidy events were irreversible. This model was able to account for the distribution of ploidal classes across nine angiosperm genera. Although this model is quite general, Meyers and Levin (2006) made a number of simplifying assumptions: 1.) congeneric diploids and higher ploids shared a single diversification and polyploidization rate, 2.) all polyploids were allopolyploids, 3.) no extinction, and 4.) deterministic evolution. Here, I have advanced our understanding of angiosperm evolution by relaxing these assumptions and analyzing the broadest phylogenetic distribution of species yet considered in a comparative study of polyploid evolutionary rates.

Fifty-five of the 60 flowering plant genera considered have ploidal distributions statistically consistent with the simple Meyers and Levin ratchet model. These genera represent 22 of the 25 families and 14 of the 17 orders contained in my full data set. I therefore conclude that this provides strong support for the sufficiency and utility of the ratchet model of polyploid evolution. Although variation in diversification rates has undoubtedly contributed to the expansion of angiosperm genera, the abundance of polyploids observed in many lineages can be explained without assuming that polyploids

are superior to diploids. This finding highlights again the importance of irreversibility in shaping evolutionary trajectories (Bull and Charnov 1985).

At first glance, my assumption about polyploid irreversibility may seem at odds with the recent suggestion that all extant angiosperms have descended from a polyploid ancestor (Soltis et al. 2009). However, this assumption need only hold over the evolutionary time-scales considered in each simulation. For example, Jiao et al. (2011) estimated that two ancestral whole genome duplications shared by all angiosperms occurred between 100 – 300 and 200 – 450 mya, whereas the relatively genera in my study have a mean age of 22 million years. Thus, my study only assumes irreversibility over a relatively short and recent period of angiosperm evolution. Furthermore, another recent study found statistical support for the irreversibility of polyploidy (Mayrose et al. 2011).

Polyploidy is infrequent, in part, because hybridization events in which unreduced gametes unite to form viable and fertile polyploids are rare (Soltis and Soltis 1999). These nascent polyploids then must overcome a substantial minority disadvantage (Levin 1975). Consequently, newly formed polyploids frequently exhibit characteristics such as high levels of self-fertilization, assortative mating, divergent habitat preferences, or a substantial fitness advantage over their progenitors (Soltis and Soltis 1999; Otto 2007). Given these restrictive conditions surrounding polyploid emergence in the presence of their lower-ploid progenitors, my model distinguishes the polyploid lineage formation process from subsequent diversification.

Importantly, my model considers only random extinction. Since I assume all genera a founded by diploids, random extinction ultimately reduces diploid diversity proportionally more than polyploidy diversity. Incorporating non-random extinction, for example, an upper bound on viable genome size, may impact the evolutionary rate

estimates, which I cannot predict *a priori*. Such extensions of my model to include more complex evolutionary processes should prove insightful.

I lacked statistical power to estimate proportion of polyploid lineages founded by autopolyploidy versus allopolyploidy. My method gauges the balance between the two forms of polyploidy primarily from observed numbers of hexaploids and other ploidal levels that cannot be created via autopolyploidy alone. However, there were too few higher ploidal species in most of my focal genera to estimate the relative importance of autopolyploidy, and it thus remains an important topic for future investigation.

My results are consistent with an emerging consensus that polyploids do not have a speciation advantage over related diploids. Wood et al. (2009) failed to detect increased diversification rates in lineages with higher ploidal levels using a method of non-nested sister group contrasts, and Mayrose et al. (2011) found evidence for a decrease in the speciation rate of recently formed polyploids. Interestingly, five of my genera were included in the Mayrose et al. (2011) analysis and, despite different data and methods, the qualitative conclusions agree. Importantly, the Mayrose et al. (2011) method considered phylogenetic relationships within genera, while mine do not.

Speculation about polyploidy advantage has been motivated, in part, by the positive relationship between species diversity and the incidence of polyploidy. I show that this relationship arises naturally from polyploidy irreversibility, and can occur even if polyploids have an evolutionary disadvantage. Concordantly, Vamosi and Dickinson reported a correlation between polyploidy incidence and species richness in Rosaceae (2006), yet found no evidence that polyploids diversified faster than their diploid counterparts. They concluded that ploidal evolution alone could account for the pattern. When Mayrose *et al*. (2010) recreated chromosome number evolution in *Helianthus*, they hypothesized that major polyploid events are followed by depressed speciation rates.

These studies have led some to conclude that polyploids should, in fact, be considered evolutionary dead-ends (Arrigo and Barker 2012).

My results suggest that a simple ratchet model for polyploid evolution can explain the within-genus distribution of ploidal levels across angiosperms. I argue that this model should serve as a null model for future studies on polyploidy and diversity. I have taken this approach in evaluating more complex evolutionary drivers, and find statistical evidence for a diploid advantage driven by the relatively frequent emergence of new tetraploid species. I conclude that the rise of polyploids and the concomitant rise of biodiversity do not require the evolutionary superiority of polyploidys. Nonetheless, polyploids are central evolutionary drivers that should not be relegated to dead-ends.

The Models, Methods, and Data

The data

My analysis was based on two empirical data sets. First, I used the observed ploidal level distributions of 60 flowering plant genera chosen haphazardly from several issues of the Missouri Botanical Garden Index to Chromosome Numbers, spanning the years 1967 to 2000 (Table 2) (Moore 1973; Goldblatt 1981; Missouri Botanical Garden 2005). I sought genera for which chromosome counts were available for large numbers of species and included at least three ploidal levels. To determine the ploidal level of a species, I followed the rationale of Grant (1981). For example, if a genus contains species with 20, 40, or 60 chromosomes species, then 2n=20 would be counted as diploid, 4n=40 tetraploid, and 6n=60 hexaploid. Where minor intraspecific variation was present and one number prevailed, that number was used. If there was extensive variation, the species was not considered.

Second, I considered estimates of the self-fertilization rate, defined as the proportion of offspring produced by self-fertilization, for 16 of the genera included in this study, as reported in (Barringer 2007). I used these data to test for a correlation between these mean self-fertilization rates and the polyploid formation rates estimated in my analysis. Importantly, the mean self-fertilization rates were not estimated as part of this study nor included as a model parameter.

Genus	Total	2n	4n	6n	8n	10n	12n	14n	16n
Acacia	1200	67	16	0	1	0	0	0	1
Aconitum	300	77	45	2	0	0	0	0	0
Agropyron	80	17	39	11	2	1	0	0	0
Agrostis	220	23	34	23	2	0	1	0	0
Allium	690	249	48	3	1	0	0	0	0
Anemone	144	53	17	4	0	0	0	0	0
Arenaria	150	45	29	6	2	0	0	0	0
Artemisia	350	72	37	12	1	0	0	0	0
Aster	250	103	38	26	9	0	3	0	0
Astragalus	1700	214	53	7	3	2	0	0	0
Atriplex	300	39	17	5	1	0	0	0	0
Bromus	100	26	30	6	7	3	0	0	0
Calamagrostis	230	36	13	1	1	0	0	0	0
Calceolaria	388	43	43	5	3	0	0	0	0
Camillia	200	79	10	15	2	0	0	0	0
Castilleja	200	65	22	4	4	0	0	0	0
Centaurea	500	137	43	5	2	0	0	0	0
Cerastium	100	49	24	5	7	0	0	0	0
Chenopodium	100	42	16	9	0	2	0	0	0
Chrysanthemum	60	36	12	5	2	2	0	0	0
Cuphea	260	65	60	17	11	0	0	0	0
Draba	300	32	10	5	9	3	0	0	1
Elymus	150	8	105	32	5	0	1	0	0
Erigeron	150	130	13	14	3	0	0	0	0
Eriospermum	100	69	10	7	1	0	0	0	0
Eryngium	240	64	14	7	4	3	1	0	0
Eupatorium	1200	75	27	6	6	0	2	0	0
Euphorbia	2000	115	57	15	2	1	2	0	0
Festuca	450	52	45	43	10	2	0	0	0

Table 2 – Ploidal level distributions.

Galium	300	71	36	4	4	2	0	0	0
Geranium	300	2	47	7	1	1	0	0	0
Gymnocalycium	65	47	13	2	0	0	0	0	0
Hibiscus	300	24	20	8	1	1	1	0	0
Hieracium	90	66	96	9	0	0	0	0	0
Leucanthemum	50	23	8	8	2	2	0	0	0
Minuartia	100	68	14	2	0	0	0	0	0
Muhlenburgia	160	44	32	4	1	0	0	0	0
Opuntia	200	50	20	11	6	1	0	0	0
Oxytropis	300	74	24	23	3	1	3	0	0
Panicum	500	73	44	14	4	0	0	0	0
Paspalum	330	46	56	19	8	0	1	0	0
Pelargonium	280	125	30	6	4	0	0	0	0
Plantago	270	58	34	1	3	0	2	1	0
Poa	200	22	26	20	11	2	0	0	0
Polygonum	120	31	30	12	3	0	0	0	0
Potentilla	500	58	61	34	13	2	0	0	0
Ranunculus	600	151	90	31	3	0	4	0	1
Rosa	125	14	13	7	1	0	0	0	0
Rubus	250	59	119	4	6	0	0	0	0
Rumex	200	31	29	9	5	0	3	0	0
Salix	400	40	12	5	1	0	0	0	0
Senecio	1250	53	180	59	24	15	2	1	2
Setaria	150	15	28	7	3	0	1	0	0
Silene	700	138	7	2	1	0	0	0	0
Solanum	1700	142	30	15	0	0	0	0	0
Solidago	110	84	14	5	0	0	1	0	0
Stachys	300	17	66	2	14	2	1	0	0
Trifolium	249	152	13	9	1	0	0	0	0
Veronica	180	56	25	8	2	1	0	0	0
Viola	400	79	25	15	2	1	0	0	0

Table 2 Continued – Ploidal level distributions.

The Polyploid Ratchet Model

The model assumes that a genus is founded by a single diploid species and then tracks the evolutionary dynamics of the genus in terms of changing numbers of species at each ploidal level. The ploidal level of each species in the genus is given with respect to the original founder. I let x_k denote the number of species of ploidal level k, and

consider only even values of k. For a species with ploidal level k, r_k denotes the within-ploidal level diversification rate, which is the rate species give rise to daughter lineages with the same ploidal level, and h_k denotes the rate at which the species gives rise to new species through successful polyploidy events. The resulting polyploids are autopolyploids with probability a and allopolyploids with probability $(1-a)$. Extinction happens at a background rate μ. Using these parameters, the net speciation rate for ploidal level k is $\lambda_k = r_k + h_k - \mu$.

In this study, I consider a simple and complex version of this model:

Simple Ratchet (two parameters): The original Meyers and Levin (2006) model where all species in a genus share a single diversification rate, r, a single polyploidization rate, h, all polyploids are allopolyploids ($a = 0$), and no extinction ($\mu = 0$).

Complex Ratchet (five parameters): Polyploid species share a single rate polyploidization $(h_p = h_4 = h_6 = h_8 = \cdots)$ and diversification rate $(r_p = r_4 = r_6 = r_8 = \cdots)$ that can be different from the diploid polyploidization ($h_d = h_2$) and diversification rates ($r_d = r_2$). The fraction of autopolyploids (a) can take on non-zero values. Extinction is set at a fixed rate for all species in a genus $(\mu = \mu_2 = \mu_4 = \mu_6 = \cdots)$.

Simulation of the Model

Here, I outline a numerical algorithm for simulating the Complex Ratchet model, henceforth called the *simulation*. This continuous-time stochastic algorithm has evolutionary parameters specific to each genus: the diversification rates (r), polyploidization rates (h), the probability that a polyploidy event yields an autopolyploid species (a), the final size of the genus (number of species = N_g), the estimated age of the

genus (T_g million years), and extinction rate (μ). Each genus is simulated separately, beginning with a single diploid species at time zero and tracking the number of species in each ploidal class as the genus diversifies through speciation and polyploidization.

In the continuous time version of the model, the times between both speciation and polyploidization events for a single lineage are distributed exponentially. The simulation has a constantly updating queue of speciation and polyploidization events, and iteratively performs the first event in the queue until the genus reaches its final size. Each event has three pieces of information: type (speciation or polyploidization), time (t_e, million years), and species (A_e). Speciation events occur as follows:

(S1) The simulation clock is updated to the time of the event (t_e).

(S2) The parent species (A) persists and a new offspring species (B) is created.

(S3) Species A is assigned a new time until speciation (σ_a). This value is a random deviate from an exponential distribution with rate r_d or r_p if A is a diploid or polyploid, respectively.

(S4) Species B is assigned a time until speciation (σ_b) and a time until polyploidization (γ_b). These values are random deviates from an exponential distribution with rates r_d or r_p and h_d or h_p, depending on whether the species is a diploid or higher ploid.

(S5) Event times σ_a, σ_b, and γ_b are inserted into the queue and the queue is sorted in ascending order, based on the timing of the event.

For polyploidization events, a random deviate from a Bernoulli distribution with probability of success (a) is drawn and the event will be an autopolyploid event if a one is selected and an allopolyploid event otherwise. If it is an allopolyploid event, a second parent species is required. The simulation checks whether another species is waiting to form an allopolyploid. If so, polyploidization occurs and if not, the parent species is

waitlisted for polyploidization (at any time, there is at most a single species waiting). During an allopolyploidization event the following operations occur:

(P1) The simulation clock is updated to the time of the event (t_e) [the latter of the two parental polyploidization times].

(P2) The parent species, A_1 and A_2 with ploidal levels k_1 and k_2, persist and new offspring, B, is created with ploidal level $k_1 + k_2$.

(P3) Species A_1 and A_2 are assigned new times until polyploidization (γ_{a1} and γ_{a2}). These times are random deviates from an exponential distribution with rate h_k.

(P4) The offspring species (B) is assigned a time until speciation (σ_b) and a time until polyploidization (γ_b), as per (S4).

(P5) The event times γ_{a1}, γ_{a2}, γ_b, and σ_b are added to the queue such that the queue remains sorted from the earliest to the latest event.

For autopolyploidy events, the daughter species will have a ploidal level double that of the parent's. The parent species will be given a new polyploidization and the daughter species polyploidization and speciation waiting-time as described above.

Extinction is modeled using an exponential waiting time with rate $\mu = 0.1$. When each new species is formed, a random deviate from an exponential distribution with rate μ is drawn and that plus the current simulation clock time becomes the extinction time for that species. Once the simulation clock reaches the extinction time, that species is removed from all queues. If a species is waiting to form an allopolyploid when extinction occurs, the daughter species are not created.

Parameter Estimation

Of the parameters included in the model (r_k,a, h_k, T_g, N_g, and μ), I used fixed values for T_g, N_g, and μ. The total genus size (N_g) and age of the genus (T_g) were taken from the literature; where no T_g was reported, I assumed that species had an age equal to the average of those with observed ages, $T_g = 22\text{my}$. Ages were taken from the Missouri Botanical Garden Index to Chromosome Numbers, spanning the years 1967 to 2000 (Moore 1973; Goldblatt 1981; Missouri Botanical Garden 2005). To estimate the remaining parameters I employed the aforementioned simulation and the model fitting procedure Approximate Bayesian Computation – Sequential Monte Carlo (ABC-SMC) (Beaumont et al. 2002; Toni et al. 2008). I used ABC-SMC because closed form expressions for the likelihood equation necessary to calculate the joint-posterior distribution of r_k,h_k, and a do not exist. The lack of a closed-form likelihood equation derives primarily from the presence of allopolyploidy in the model.

I now briefly describe the ABC-SMC parameter estimation procedure (see Toni et al. (2008) for a complete discussion of the methodology). If a model M with key parameter θ, that has a prior distribution $\pi(\theta)$, generates some data D then the posterior distribution for θ is given by:

$$f(\theta \mid D) = \frac{\Pr(D \mid \theta)\pi(\theta)}{\int \Pr(D \mid \theta)\pi(\theta)d\theta}.$$

Since it is not possible to calculate the likelihood $\Pr(D \mid \theta)$ directly for my model, I implement the following approximate method for estimating $f(\theta \mid D)$:

Select a vector of decreasing acceptance levels, $E = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$, such that $\varepsilon_1 > \varepsilon_2 > ... > \varepsilon_n$.

Select a final number of acceptances, S, for each level in vector *E*.

Generate a parameter set Θ by selecting a random deviate from the joint prior distribution $\Theta = \pi(\theta_{h_k}) \circ \pi(\theta_{r_k}) \circ \pi(\theta_a)$.

Simulate a data set D_r with parameter set Θ.

Calculate the distance between the observed data, D_{obs}, and the simulated data, D_r, using a set of pre-specified distance metrics $\delta(D_{obs}, D_r)$. In this case, I calculated the Euclidean distance on nine summary statistics that have been transformed using Partial Least Squares regression (summary statistics: total size, proportion diploid, proportion tetraploid, proportion tetraploid or greater, proportion hexaploid or greater, highest observed ploidal level, proportion highest ploidal level, average ploidal level, variance in ploidal level).

Accept Θ if $\delta(D_{obs}, D_r) < \varepsilon_i$.

Repeat steps 3-5 until S sets of parameters are accepted, each time recording the parameter set and the corresponding distance, $\delta(D_o, D_r)$.

Once S sets of parameters are accepted, update the prior distributions to be equivalent to the distribution of accepted parameters and add noise to the distributions using a uniform kernel.

Update the tolerance to ε_{i+1} and repeat steps 2-6.

Once S sets of parameters are accepted at tolerance ε_n, the algorithm stops and the accepted parameter sets are taken as the joint posterior distribution over the parameters Θ.

Comparative Analyses

To assess whether there is evidence of a polyploid advantage or disadvantage in terms of relative evolutionary rates across my study genera, I considered parameter estimates under the Complex Ratchet model. To evaluate the fit of Simple Ratchet

model, I used the goodness-of-fit test described below. I controlled for phylogenetic non-independence using a method proposed by Lajeunesse (2009) for use in meta-analyses. First, I constructed a phylogeny of angiosperm genera using data from Jansen et al. (2007). To fill in taxa not represented in the Jansen et al. (2007) data set, I used family level relationships while maintaining the original topology. I then calculated the difference between the diploid and polyploid evolutionary rates, for example the diversification rate difference for each genus would, $r_d - r_p$, and the pooled variance $\sigma^2(r_d - r_p)$. Using a random effects model that incorporates a variance/covariance matrix weighting each estimate based on how much of the variance/covariance structure can be explained by phylogeny, I determined the phylogenetically independent effect size and 95% confidence interval. Finally, I determined whether each difference in evolutionary rate was significantly different from zero. These calculations were performed using the software package PHYLOMETA (Lajeunesse 2011).

Polyploidy and Species Richness

To investigate the contribution of polyploidy to within-genus species richness, I compared two models: one where $h = 0$ and another where $h > 0$. Under each model, I simulate forward in time, recording the absolute number of species at time T_G. The net speciation rates of diploid and higher ploids were set to be equal in both models ($\lambda_p = \lambda_d$), allowing me to focus on the relative contribution of polyploid formation to within-genus species richness.

Goodness-of-Fit Test

I used a goodness-of-fit test to determine whether the Simple Ratchet model could generate distributions of ploidal levels statistically similar to the empirical distributions, for each of the 60 genera included in this study. Because the number of observations in individual ploidal classes is often too low to perform a standard Chi-Squared test, null distributions must be generated via Monte Carlo simulation. Briefly, the estimated parameters values are used to simulate 1000 ploidal level distributions for each genus. For every simulated distribution I calculate a Chi-Square statistic,

$$\chi_g^2 = \sum_{k=2}^{16} \frac{(O_k - E_k)^2}{E_k}$$

where O_k is the simulated or observed number of species in ploidal class k and E_k is the actual number or expected. By aggregating each of the 1000 Chi-Square statistics, I can calculate a *P-value* that is the proportion of this distribution that is greater than or equal to the Chi-Square statistic calculated using the Simple Ratchet model.

Chapter 3: Optimizing Provider Recruitment for Surveillance Networks[3]

Abstract

The increasingly complex and rapid transmission dynamics of many infectious diseases necessitates the use of new, more advanced methods for surveillance, early detection, and decision-making. Here, I demonstrate that a new method for optimizing surveillance networks can improve the quality of epidemiological information produced by typical provider-based networks. Using past surveillance and Internet search data, it determines the precise locations where providers should be enrolled. When applied to redesigning the provider-based, influenza-like-illness surveillance network (ILINet) for the state of Texas, the method identifies networks that are expected to significantly outperform the existing network with far fewer providers. This optimized network avoids informational redundancies and is thereby more effective than networks designed by conventional methods and a recently published algorithm based on maximizing population coverage. I show further that Google Flu Trends data, when incorporated into a network as a virtual provider, can enhance but not replace traditional surveillance methods.

[3] Considerable portions of this chapter were published as Scarpino SV, Dimitrov NB, and Meyers LA. 2012. Optimizing Provider Recruitment for Influenza Surveillance Networks. PLoS Comput Biol 8(4): e1002472. **Contributions** - Conceived and designed the experiments: SVS NBD LAM. Performed the experiments: SVS NBD. Analyzed the data: SVS NBD. Contributed reagents/materials/analysis tools: SVS NBD. Wrote the paper: SVS NBD LAM.

Introduction

Since the Spanish Flu Pandemic of 1918–1919, the global public health community has made great strides towards the effective surveillance of infectious diseases. However, modern travel patterns, heterogeneity in human population densities, proximity to wildlife populations, and variable immunity interact to drive increasingly complex patterns of disease transmission and emergence. As a result, there is an increasing need for effective, evidence-based surveillance, early detection, and decision-making methods (Brownstein et al. 2008; Khan et al. 2010; Mnatsakanyan et al. 2011). This need was clearly articulated in 2009 by a directive from the Department of Homeland Security and the Centers for Disease Control and Prevention to develop a nationwide, real-time public health surveillance network (Bush 2007; CDC 2010).

The U.S. Outpatient Influenza-Like Illness Surveillance Network (ILINet) gathers data from thousands of healthcare providers across all fifty states. Throughout influenza season (CDC mandating reporting during weeks 40 – 20, which is approximately October through mid-May), participating providers are asked to report weekly the number of cases of influenza-like illness treated and total number of patients seen, by age group. Cases qualify as ILI if they manifest fever in excess of 100◦F along with a cough and/or a sore throat, without another known cause. Although the CDC receives reports of approximately 16 million patient visits per year, many of the reports may use a loose application of the ILI case definition and/or may simply be inaccurate. The data are used in conjunction with other sources of laboratory, hospitalization and mortality data to monitor regional and national influenza activity and associated mortality. Similar national surveillance networks are in place in 11 EU countries and elsewhere around the globe (Ordobas et al. 1995; Carrat et al. 1998; Clothier et al. 2006; Deckers et al. 2006).

Each US state is responsible for recruiting and managing ILINet providers. The CDC advises states to recruit one regularly reporting sentinel provider per 250,000 residents, with a state-wide minimum of 10 sentinel providers. Since 2003, the Texas Department of State Health Services (DSHS) has enrolled a total of 300 volunteer providers. Participating providers regularly drop out of the network; Texas DSHS aims to maintain approximately 200 active participants through year-round recruitment of providers in heavily populated areas (cities with populations of at least 100,000). DSHS also permits other (non-targeted) providers of family medicine, internal medicine, pediatrics, university student health services, emergency medicine, infectious disease, OB/GYN and urgent care to participate in the network. During the 2009 – 2010 influenza season, the Texas ILINet included 205 providers with approximately 50% reporting most weeks of the influenza season.

A number of statistical studies have demonstrated that ILI surveillance data is adequate for characterizing past influenza epidemics, monitoring populations for abnormal influenza activity, and forecasting the onsets and peaks of local influenza epidemics (Quenel and Dab 1998; Fleming et al. 1999; Viboud et al. 2003; Rath et al. 2003; Cowling et al. 2006; Jiang et al. 2009; Yang et al. 2009). However, the surveillance networks are often limited by non-representative samples (Polgreen et al. 2009), inaccurate and variable reporting (Quenel and Dab 1998; Fleming et al. 1999; Yang et al. 2009), and low reporting rates (Clothier et al. 2006). Some of these studies have yielded specific recommendations for improving the performance of the surveillance network, for example, inclusion of particular categories of hospitals in China (Yang et al. 2009), preference for general practitioners over pediatricians in Paris, France (Quenel and Dab 1998), and a general guideline to target practices with high reporting rates and high numbers of patient visits (per capita) (Clothier et al. 2006). Polgreen et al. (2009) recently

described a computational method for selecting ILINet providers so as to maximize coverage, that is, the number of people living within a specified distance of a provider (Polgreen et al. 2009). They applied the approach to optimizing the placement of the 22 providers in the Iowa ILINet. While their algorithm ensures maximum coverage, it is not clear that maximum coverage is, in general, the most appropriate criterion for building a statistically informative ILINet.

In 2008, Google.org launched Google Flu Trends, a website that translates the daily number of Googles search terms associated with signs, symptoms, and treatment for acute respiratory infections into an estimate of the number of ILI patients per 100, 000 people. It was shown that Google Flu Trends reliably estimates national influenza activity in the US (Ginsberg et al. 2008), the state of Utah (Ginsberg et al. 2008), and in some European countries (Valdivia et al. 2010), but it provided imperfect data regarding the 2009 H1N1 pandemic in New Zealand (Wilson et al. 2009). I assessed the correlation between Google Flu Trends for Texas and Texas' ILINet data and found a correlation of 0.87, similar to those presented in Ginsberg et al. 2008 (Ginsberg et al. 2008). The Google Flu Trends website includes ILI-related search activity down to the level of cities (in beta version as of November 2011). Thus, Google Flu Trends may serve as a valuable resource for influenza detection and forecasting if effectively integrated with public health data such as those coming from state ILINets.

Here, I present an evaluation of the Texas Influenza-Like-Illness Surveillance Network (ILINet), in terms of its ability to forecast statewide hospitalizations due to influenza (ICD9 487 and 488) and unspecified pneumonia (ICD9 486). Although I henceforth refer to this subset of hospitalizations as influenza-like hospitalizations, I emphasize that these data do not perfectly reflect influenza-related hospitalizations: some unrelated pneumonias may be classified under ICD9 486, and

some influenza cases may not be correctly diagnosed and/or recorded as influenza. Nonetheless, this subset of hospitalizations likely includes a large fraction of hospitalized influenza cases and exhibits strong seasonal dynamics that mirror ILINet trends. The inclusion of all three ICD9 codes was suggested by health officials at Texas DSHS who seek to use ILINet to ascertain seasonal influenza-related hospitalization rates throughout the state (Texas DSHS contract numbers 2009–032591 and 2011–037903). Hospitalizations associated with these three codes in Texas accounted for between 20 and 35% of all hospitalizations due to infections and roughly 9.5 billion dollars of hospitalization payments in 2008.

Using almost a decade of state-level ILINet and hospitalization data, I find that the existing network performs reasonably well in its ability to predict influenza-like hospitalizations. However, smaller, more carefully chosen sets of providers should yield higher quality surveillance data, which can be further enhanced with the integration of state-level Google Flu Trends data. For this analysis, I adapted a new, computationally tractable, multilinear regression approach to solving complex subset selection problems. The details of this method are presented below and can be tailored to meet a broad range of surveillance objectives.

Results

Using a submodular ILINet optimization algorithm, I investigate two scenarios for improving the Texas ILINet: designing a network from scratch and augmenting the existing network. I then evaluate the utility of incorporating Google Flu Trends as a virtual provider into an existing ILINet.

Designing a new ILINet

To construct new sentinel surveillance networks, I choose individual providers sequentially from a pool of approximately 2000 mock providers, one for each zip code in Texas, until I reach 200 total providers. At each step, the provider that most improves the quality of the epidemiological information produced by the network is added to the network. I optimize and evaluate the networks in terms of the time-lagged statistical correlation between aggregated ILINet provider reports (simulated by the model) and actual statewide influenza-like hospitalizations. Specifically, for each candidate network, I perform a least squares multilinear regression from the simulated ILINet time series to the actual Texas hospitalization time series, and use the coefficient of determination, R^2, as the indicator of ILINet performance. Henceforth, I will refer to these models as ILINet regression models.

I compare the networks generated by this method to networks generated by two naive models and a published computational method (Polgreen et al. 2009) (Figure 11). Random selection models an open call for providers and entails selecting providers randomly with probabilities proportional to their zip code's population; Greedy selection prioritizes providers strictly by the population density of their zip code. Submodular optimization significantly outperforms these naive methods, particularly for small networks, with Random selection producing slightly more informative networks than Greedy selection. The Geographic optimization method of Polgreen et al. (2009) selects providers to maximize the number of people that live within a specified "coverage distance" of a provider. Submodular optimization consistently produces more informative networks than this method at a 20-mile coverage distance (Figure 11) (5, 10, and 25 mile coverage distances perform worse, not shown). To visualize the relative performance of several of these networks, I compared their estimates of influenza-like hospitalizations

(by applying each ILINet regression model to simulated ILINet report data) to the true state-wide hospitalization data (Figure 12). The time series estimated by a network designed using submodular optimization more closely and smoothly matches true hospitalizations than both the actual 2008 Texas ILINet and a network designed using geographic optimization (each with 82 providers).

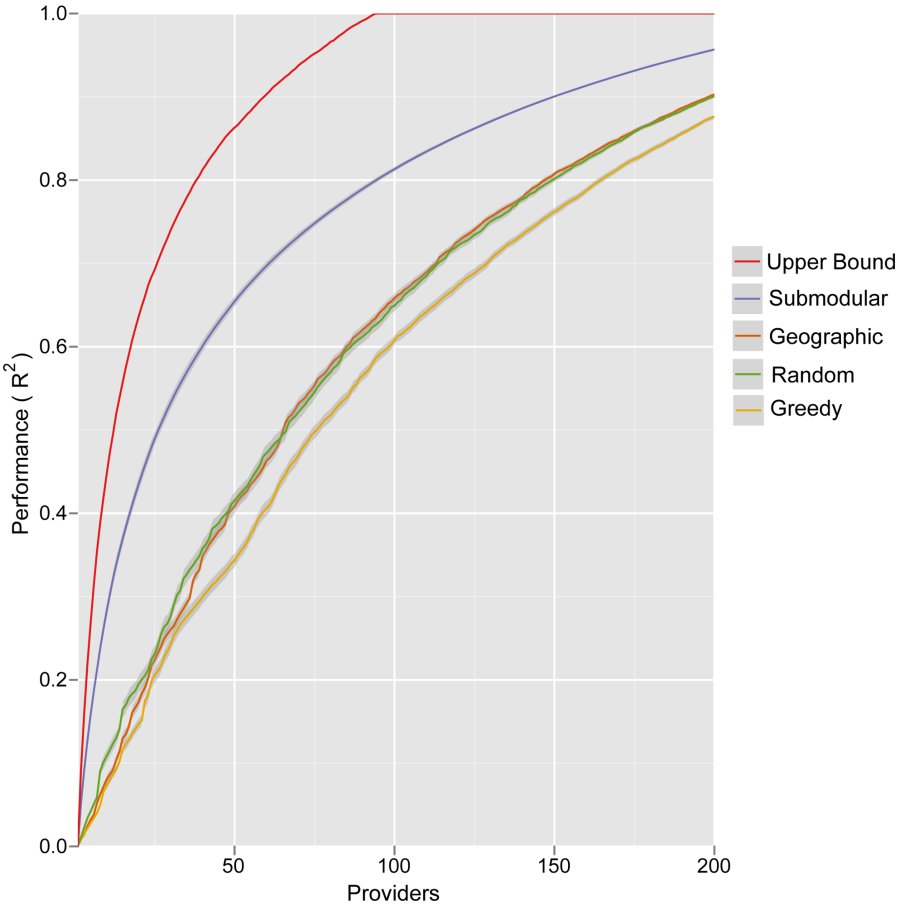

Figure 11 – Expected performance of optimized ILINets.

Four different methods were used to design Texas ILINets that effectively predict state-wide influenza hospitalizations. Submodular optimization (Submodular) outperforms

random selection proportional to population density (Random), greedy selection strictly in order of population density (Greedy), and geographic optimization to maximize the number of people that live within 20 miles of a provider (Polgreen et al. 2009) (Geographic). The theoretical upper bound for performance (Upper Bound) gives the maximum R^2 possible for a network designed by an exhaustive evaluation of all possible networks of a given size. For each network of each size, the following procedure was repeated 100 times: randomly sample a set of reporting profiles, one for each provider in the network; simulate an ILI time series for each provider in the network; perform an ordinary least squares multilinear regression from the simulated provider reports to the actual statewide influenza hospitalization data. The lines indicate the mean of the resulting R^2 values, and the error bands indicate the middle 90% of resulting R^2 values, reflecting variation stemming from inconsistent provider reporting and informational noise.

The submodular optimization algorithm is not guaranteed to find the highest performing provider network, and an exhaustive search for the optimal 200provider network from the pool of 2000 providers is computationally intractable. However, the submodular property of the objective function allows me to compute an upper bound on the performance of the optimal network, without knowing its actual composition (Figure 11). The performance gap between the theoretical upper bound and the optimized networks may indicate that the upper bound is loose (higher than the performance of the true optimal network) and/or the existence of better networks that might be found using more powerful optimization methods.

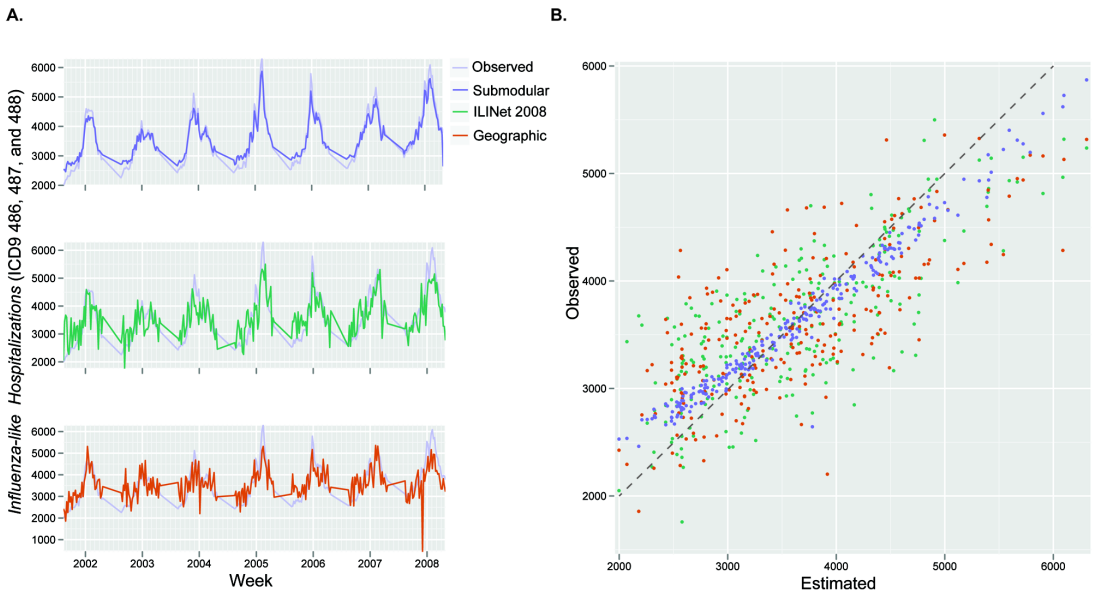

Figure 12 – Comparing ILINet estimates to actual state-wide influenza hospitalizations.

Statewide hospitalizations are estimated using data from three ILINets: the 2008 Texas ILINet (ILINet 2008), which consisted of 82 providers, and ILINets of the same size that were designed using submodular optimization (Submodular) and maximum coverage optimization with a 20 mile coverage distance (Geographic). (a) The estimates from each network are compared to actual Texas state-wide influenza hospital discharges from 2001–2008 (Observed). (b) The submodular ILINet yields estimates that are consistently closer to observed values than the other two ILINets. For each of the three networks, the following procedure was repeated 100 times: randomly sample a set of reporting profiles, one for each provider in the network; simulate an ILI time series for each provider in the network; perform an ordinary least squares multilinear regression from the simulated provider reports to the actual Texas influenza hospitalization data; and apply resulting regression model to the simulated provider time series data to produce estimates of statewide hospitalizations. The figures are based on averages across the 100 estimated hospitalization time series for each ILINet.

The networks selected by submodular optimization reveal some unexpected design principles. Most of the Texas population resides in Houston and the "I-35 corridor" – a North-South transportation corridor spanning San Antonio, Austin, and Dallas (Figure 13a). The first ten provider locations selected by submodular optimization are spread throughout the eastern half of the state (Figure 14a, pink circles). While most of the providers are concentrated closer to Texas' population belt, only two are actually located within Texas' major population centers (in this case, College Station).

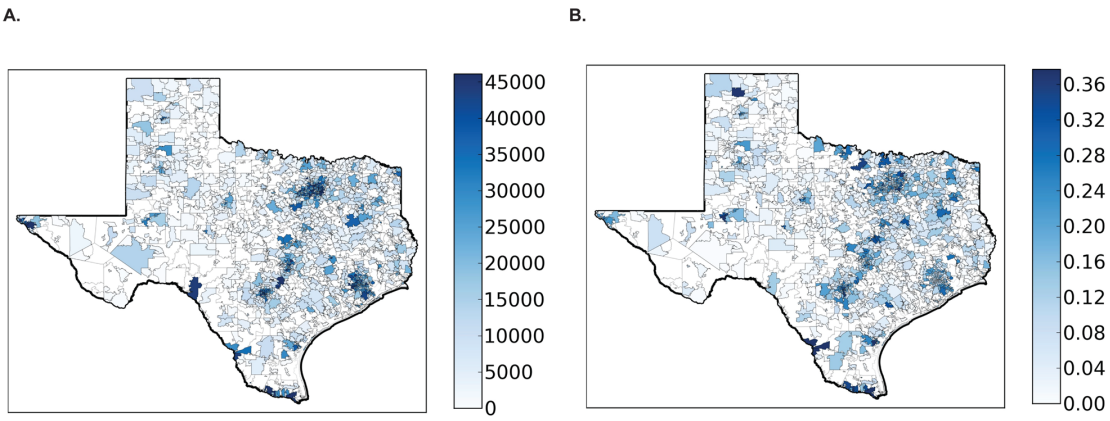

Figure 13 – Statewide influenza activity mirrors population distribution.

(a) Shading indicates zip code level population sizes, as reported in the 2000 census. (b) Major populations centers exhibit covariation in influenza activity. I performed a principal component analysis (PCA) on the centered hospitalization time series of all zip codes and calculated the time series of the first principal component. Zip codes are shaded according to the R^2 obtained from a regression of the first principal component time series to the influenza hospitalization time series for the zip code. Dark shading indicates high synchrony between influenza activity in the zip code and the first principal

component. The correspondence between darkly shaded zip codes in (a) and (b) results from the high degree of synchrony in influenza activity between highly populated zip codes in Texas.

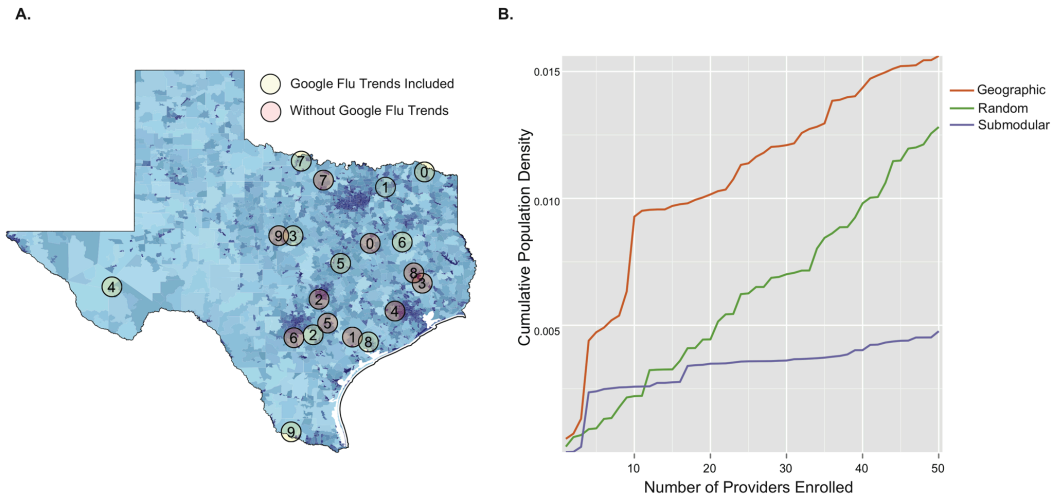

Figure 14 – Location and population coverage of optimized ILINets.

(a) Shading indicates zip code level population sizes, as reported in the 2000 census. Circles indicate the location (zip code) of the first ten providers selected when Google Flu Trends is included as a provider (green) and when it is not (pink). Numbers indicate selection order, with zero being the first provider selected and nine the tenth provider selected. (b) The cumulative population densities covered increase as each ILINet grows. Cumulative density is estimated by dividing total population of all provider zip codes by total area of all provider zip codes. While ILINets designed using the geographic (orange) and random (green) methods primarily target zip codes with high population densities, submodular optimization (purple) targets zip codes that provide maximal information, regardless of population density. All three networks cover approximately the same total number of people.

The submodular networks are qualitatively different from the networks created by the other algorithms considered, which focus providers within the major population centers (Figure 14b). The higher performance of the submodular ILINets suggest that over-concentration of providers in major population centers is unnecessary. Influenza levels in the major population centers are strongly correlated (Figure 13b). Thus, ILINet information from San Antonio, for example, will also be indicative of influenza levels in Austin and Dallas. This synchrony probably arises, in part, from extensive travel between the major Texas population centers.

Subsampling and augmenting an ILINet.

Using submodular optimization, I augment the 2008 Texas ILINet by first subsampling from the 82 enrolled providers and then adding up to 40 new providers. When subsampling, performance does not reach a maximum until all 82 providers are included in the network (Figure 15), indicating that each provider adds predictive value to the network. However, the theoretical upper bound plateaus around 40 providers, suggesting that smaller (more optimally chosen) networks of equal predictive value may exist. During the second stage, 40 additional providers improve the R^2 objective by 33%. Most of these providers are located in relatively remote areas of the state.

I also considered inclusion of Internet trend data sources as virtual providers, specifically, the freely available Google Flu Trends data for the state of Texas (Googleorg 2003). Google Flu Trends alone is able to explain about 60% of the variation in state-wide hospitalizations; it outperforms the 2008 Texas ILINet and matches the performance of a network with 44 traditional providers constructed from scratch using submodular optimization (Figure 16). However, the best networks include both

traditional providers and Google Flu Trends. For example, by adding 50 providers to Google Flu Trends using submodular optimization, I improve the R2 objective by a third and halve the optimality gap (from a trivial upper bound of one). The additional providers are located in non-urban areas (Figure 14a, green circles) distinct from those selected when Google Flu Trends is not allowed as a provider.

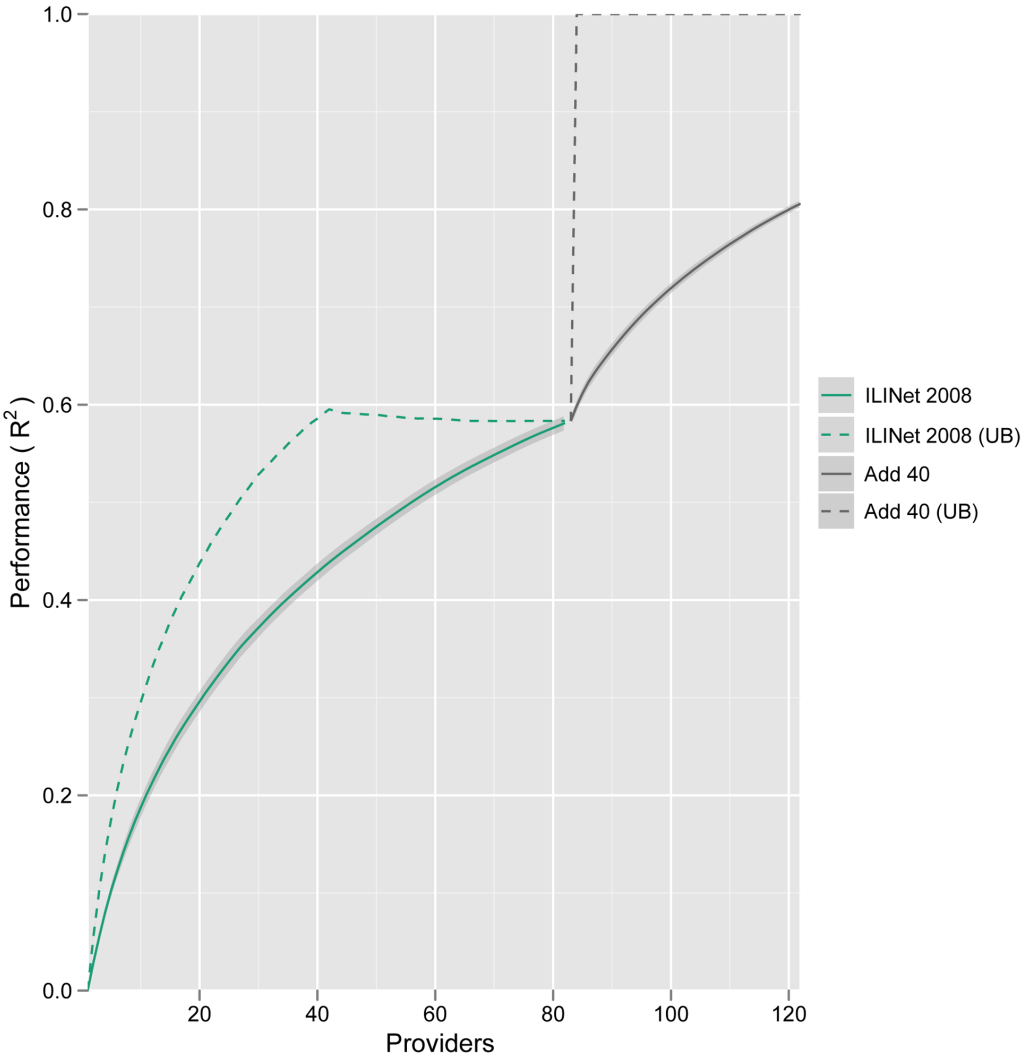

Figure 15 – Augmenting an existing ILINet.

This compares theoretical upper bounds (dashed lines) to the performance of a submodular optimized ILINet built by first subsampling the 82 zip codes of providers actually enrolled in Texas' 2008 ILINet (green) and then adding 40 additional providers from elsewhere in the state (gray). The error bands indicate the middle 90% of resulting R^2 values, and reflect variation stemming from inconsistent provider reporting rates and informational noise.

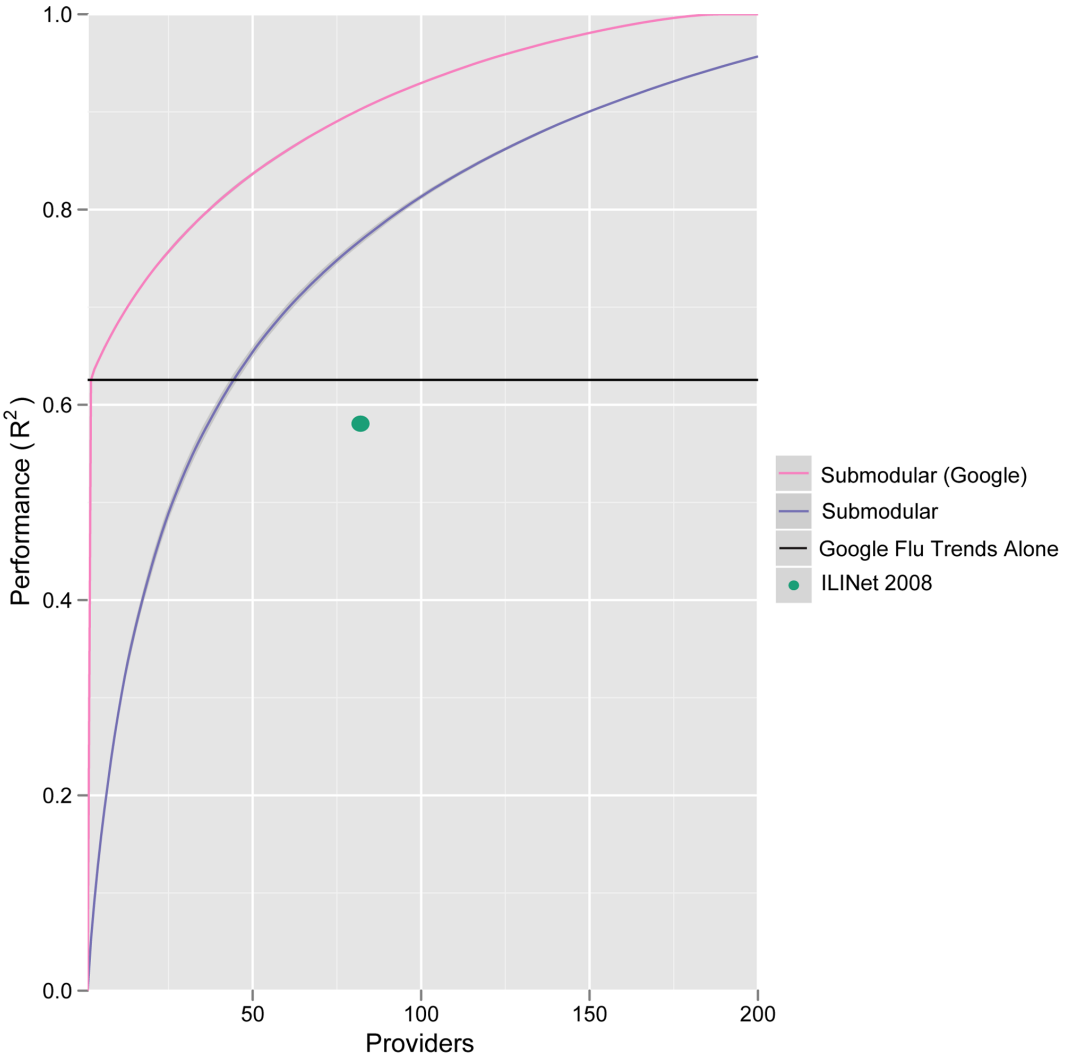

Figure 16 – Google Flu Trends as a virtual ILINet provider.

When state-level Google Flu Trends is treated as a possible provider, submodular optimization choses it as the first (most informative) provider for the Texas ILINet, and results in a high performing network (pink line). Alone (black line), the Google Flu Trends provider performs as well as a traditional submodular optimized network (blue line) containing 44 providers (intersection of black and purple lines) and outperforms the actual 2008 Texas ILINet (green dot).

Out-of-sample validation.

To further validate the methodology, I simulated the real-world scenario in which historical data are used to design an ILINet and build forecasting models, and then current ILINet reports are used to make forecasts. Specifically, I used 2001 – 2007 data to design ILINets and estimate multilinear regression models relating influenza-like hospitalizations to mock provider reports, and then used 2008 data to test the models' ability to forecast influenza-like hospitalizations. For networks with fewer than 150 providers, the ILINets designed using submodular optimization consistently outperform ILINets designed using the other three strategies (Figure 17). Above 100 providers, the predictive performance of the submodular optimization ILINet begins to decline with additional providers. As the number of providers approaches 222 (the number of weeks in the training period), the estimated prediction models become overfit to the 2001–2007 period. Thus, the slightly increased performance of the Random method over the submodular optimization after 175 providers is spurious. For the R^2 values presented in Figure 17, the effect of noise and variable reporting are integrated out when calculating the expected provider reports.

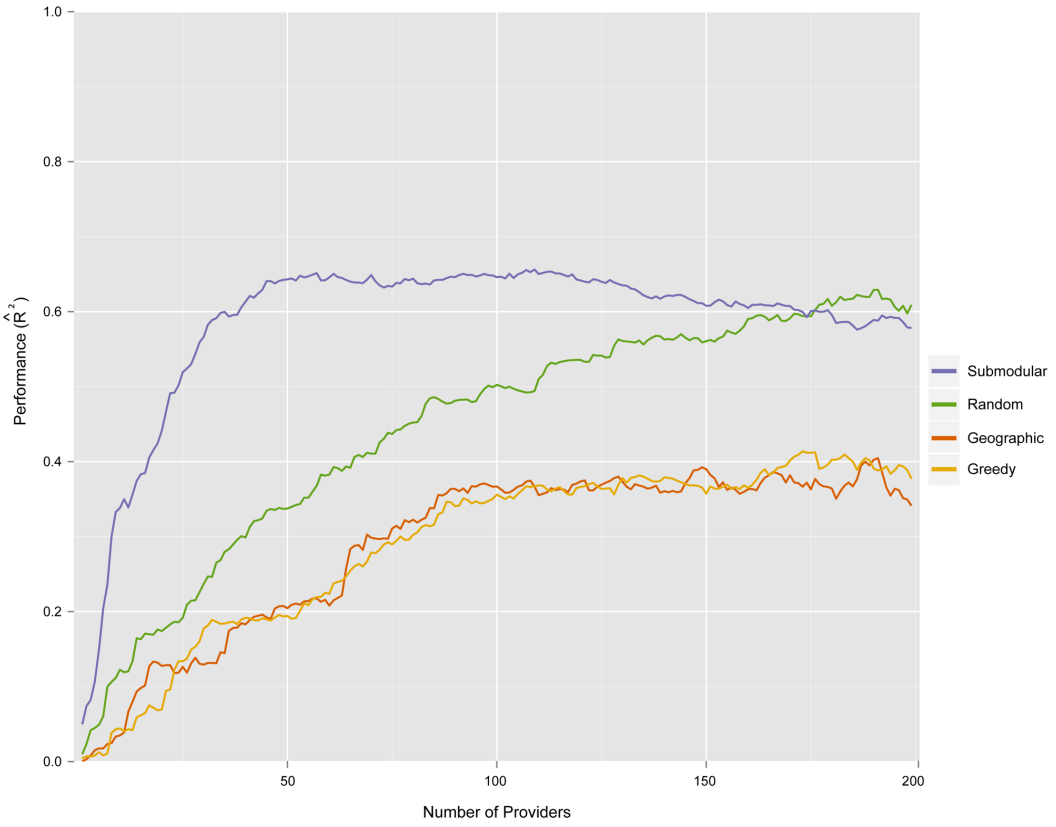

Figure 17 – Predictive performance of ILINets.

Data from the 2001–2007 period were used to design ILINets and estimate multilinear regression prediction models. The predictive performance of the ILINets (y-axis) is based on a comparison between the models' predictions for 2008 hospitalizations (from mock provider reports) and actual 2008 hospitalization data. For almost all network sizes, Submodular optimization (Submodular) outperforms random selection proportional to population density (Random), greedy selection strictly in order of population density (Greedy), and geographic optimization to maximize the number of people that live within 20 miles of a provider (Polgreen et al. 2009) (Geographic). The leveling-off of performance around 100 providers is likely a result of over-fitting, given that there were only 222 historical time-points used to estimate the original model.

Discussion

Since the mid twentieth century, influenza surveillance has been recognized as an increasingly complex problem of global concern (Langmuir and Housworth 1969). However, the majority of statistical research has focused on the analysis of surveillance data rather than the data collection itself, with a few notable exceptions (Polgreen et al. 2009; Yang et al. 2009). High quality data is essential for effectively monitoring seasonal dynamics, detecting anomalies, such as emerging pandemic strains, and implementing effective time-sensitive control measures. Using a new method for optimizing provider-based surveillance systems, I have shown that the Texas state ILINet would benefit from the inclusion of a few strategically selected providers and the use of Internet data streams.

This method works by iteratively selecting providers that contribute the most information about influenza-like hospitalizations. I quantified the performance of various ILINets using the coefficient of determination (R^2) resulting from a multilinear regression between each provider's time series and state- wide influenza-like hospitalizations. Importantly, these simulated providers have reporting rates and error distributions estimated from actual ILINet providers in Texas. The result is a prioritized list of zip codes for inclusion in an ILINet that can be used for future ILINet recruiting. Although this analysis was specifically motivated by the Texas DSHS interest in predicting hospitalizations with ICD9 codes 486, 487, and 488, the method can be readily extended to design a network for any disease or influenza definition with the appropriate historical data. In general, the method requires both historical provider reports and historical time series of the prediction target. However, if one has reasonable estimates of provider reporting rates and informational noise from another source (e.g., estimates from

a surveillance network in another region or for another disease), then historical provider reports are not necessary.

ILINet provider reports do not necessarily reflect true influenza activity. Rather they are supposed to indicate the number of patients that meet the clinical ILI case definition, which results in a substantial number of false positives (reported non-influenza cases) and false negatives (missed cases of influenza) (Monto et al. 2000). The case definition for ILI is often loosely applied, further confounding the relationship between these measures and true influenza. Similarly, the ICD9 codes used in this analysis do not correspond perfectly to influenza hospitalizations: some influenza cases will fail to be classified under those codes, and some non-influenza cases will be. Nonetheless, public health agencies are interested in monitoring and forecasting the large numbers of costly hospitalizations associated with these codes. I find that ILINet surveillance data correlates strongly with this set of influenza-like hospitalizations, and that the networks can be designed to be even more informative.

Although I provide only a single example here, this optimization method can be readily applied to designing surveillance networks for a wide range of diseases on any geographic scale, provided historical data are available and the goals of the surveillance network can be quantified. For example, surveillance networks could be designed to detect emerging strains of influenza on a global scale, monitor influenza in countries without surveillance networks, or track other infectious diseases such as malaria, whooping cough, or tuberculosis or non-infectious diseases and chronic conditions such as asthma, diabetes, cancer or obesity that exhibit heterogeneity in space, time or by population subgroup. As I have shown with Google Flu Trends, this method can be leveraged to evaluate the potential utility of incorporating other Internet trend data

mined from search, social media, and online commerce platforms into traditional surveillance systems.

While optimized networks meet their specified goals, they may suffer from over optimization and be unable to provide valuable information for other diseases or even for the focal disease during atypical situations. For example, a surveillance network designed for detecting the early emergence of pandemic influenza may look very different from one optimized to monitor seasonal influenza. Furthermore, an ILINet optimized to predict influenza-like hospitalizations in a specific socio-economic group, geographic region, or race/ethnicity may look very different from an ILINet optimized to predict state-wide hospitalizations. When optimizing networks, it is thus important to carefully consider the full range of possible applications of the network and integrate diverse objectives into the optimization analysis.

The optimized Texas ILINets described above exhibit much less redundancy (geographic overlap in providers) than the actual Texas ILINet. Whereas CDC guidelines have led Texas DSHS to focus the majority of recruitment on high population centers, the optimizer only sparsely covered the major urban areas because of their synchrony in influenza activity. This is an important distinction between submodular optimization and the other methods considered (Geographic, Random and Greedy). The submodular method does not track population density and instead adds providers who contribute the most marginal information to the network. Consequently, it places far more providers in rural areas than the other methods (Figure 14b). There can be substantial year-to-year variation in spatial synchrony for seasonal influenza, driven by the predominant influenza strains and commuter traffic between population centers (Viboud et al. 2006). As long as the historical data used during optimization reflect this stochasticity, the resulting networks will be robust. However, synchrony by geography

and population density does not occur for all diseases including emerging pandemic influenza (Viboud et al. 2006); thus the relatively sparse networks designed for forecasting seasonal influenza hospitalizations may not be appropriate for other surveillance objectives, like detecting emerging pandemic strains or other rare events. For example, a recent study of influenza surveillance in Beijing, PRC suggested that large hospitals provided the best surveillance information for seasonal influenza, while smaller provincial hospitals were more useful for monitoring H5N1 (Yang et al. 2009).

Although this method outperforms the Maximal Coverage Method (MCM), referred to as Geographic, proposed by Polgreen et al. (2009), there are several caveats. First, population densities and travel patterns within Texas are highly non-uniform. The two methods might perform similarly for regions with greater spatial uniformity. Second, the submodular method is data intensive, requiring historical surveillance data that may not be available, for example, in developing nations, whereas the population density data required for MCM is widely available. However, the type of data used in this study is readily available to most state public health agencies in the United States. For example, the CDC's Influenza Hospitalization Network (FluSurv-NET) collects weekly reports on laboratory confirmed influenza-related hospitalizations in fourteen states. In addition, alternative internet-based data sources like Google Flu Trends are becoming available. Third, as discussed above, the networks are optimized towards specific goals and may thus have no expected level of performance for alternate surveillance goals. Important future research should focus on designing networks able to perform well under a range of surveillance goals. Fourth, neither ILINet data nor influenza-like hospitalizations correspond perfectly to actual influenza activity. One could instead optimize ILINets using historical time series of laboratory-confirmed cases of influenza. Although some provider locations and the estimated regression models may

change, I conjecture that the general geospatial distribution of providers will not change significantly. Fourth, I followed Polgreen et al. (2009)'s use of Euclidean distances. However, travel distance is known to correlate more strongly with influenza transmission than Euclidean distance (Viboud et al. 2006), and thus alternative distance metrics might improve the performance of the MCM method. Finally, while submodular optimization generally outperforms the other design methods in out-of-sample prediction of influenza-like hospitalizations, it suffers from overfitting when the number of providers in the network approaches the number of data points in the historical time series.

The impressive performance of Google Flu Trends leads me to question the role of traditional methods, such as provider-based surveillance networks, in next generation disease surveillance systems. While Texas Google Flu Trends alone providers almost as much information about state-wide influenza hospital discharges as the entire 2008 Texas ILINet, an optimized ILINet of the same size contains 33% more information than Google Flu Trends alone. Adding Google Flu Trends to this optimized network as a virtual provider increases its performance by an additional 12.5%. Internet driven data streams, like Google Flu Trends, may have age and socio-economic biases that over-represent certain groups, a possible explanation for the difference in providers selected when Google Flu Trends is included, Figure 14a. Given the relatively low cost of voluntary provider surveillance networks, synergistic approaches that combine data from conventional and Internet sources offer a promising path forward for public health surveillance.

The Models, Methods, and Data

The data

The Texas Department of State Health Services (DSHS) provided (1) ILINet data containing weekly records from 2001−2010 reporting the number of patients with influenza-like-illness and the total number of patients seen by each provider in the network, and (2) individual discharge records for every hospital in Texas from 2001−2007 (excluding hospitals in counties with less than 35, 000 inhabitants, in counties with less than 100 total hospital beds, or those hospitals that do not seek insurance payment or government reimbursement). I classified all hospital discharges containing ICD9 codes of 486, 487, or 488 as influenza-related. Google Flu Trends data was downloaded from the Google Flu Trends site (Googleorg 2003) and contains estimates of ILI cases per 100,000 physician visits determined using Google searches (Ginsberg et al. 2008). Data on population size and density was obtained from the 2000 census (Censusgov 2002).

Provider Reporting Model

The first step in the ILINet optimization is to build a data-driven model reflecting actual provider reporting rates and informational noise, that is, inconsistencies between provider reports and true local influenza prevalence.

I model reporting as a Markov process, where each provider is in a "reporting" or "non-reporting" state. A provider in the reporting state enters weekly reports, while a provider in the non-reporting state does not enter reports. At the end of each week, providers independently transition between the reporting and non-reporting states. Such a Markov process model allows for streaks of reporting and streaks of non- reporting for each provider, which is typical for ILINet providers. I estimate transition probabilities

between states from actual ILINet provider report data. For each provider, the transition probability from reporting to non-reporting is estimated by dividing the number of times the transition occurred by the number of times any transition out of reporting is observed. The probabilities of remaining in the current reporting state and transitioning from non-reporting to reporting are estimated similarly.

I model noise in reports using a standard regression noise model of the form

$$(1)\ \text{Provider-repot}(i) = c_0 + c_1 PercentILI(i) + N(0,\sigma^2),$$

where Provider-report(i) denotes the number of ILI cases reported by the provider in week i; *PercentILI*(i) denotes the estimated prevalence of ILI in the provider's zip code in week i; c_0 and c_1 are regression constants fixed for the provider; and N(0,σ2) is a normally distributed noise term with variance σ2 also fixed for the provider. For existing providers, I use empirical time series (their past ILINet reporting data matched with local ILI prevalence, described below) to estimate the constants using least squares linear regression. This noise model has the intuitive interpretation that each provider's reports are a noisy reading of the percent of the population with ILI in the provider's zip code.

I use the Texas hospital discharge data to estimate the local ILI prevalences (Percent-ILI(i)) for each zip code. Given an estimate of the influenza hospitalization rate (Thompson et al. 2004) and assuming that each individual with ILI is hospitalized independently, I can obtain a distribution for the number of influenza-related hospitalizations in a zip code, given the number of ILI cases in the zip code. Using Bayes rule, a uniform prior, and the real number of influenza-related hospitalizations (from the hospital discharge data), I derive distributions for the number of ILI cases for each zip code and each week. I then set Percent-ILI(i) for each zip code equal to the mean of the

distribution of ILI cases in that zip code for week *i*, divided by the population of the zip code.

Generating Pools of Mock Providers

The second step in the ILINet optimization is to generate a pool of mock providers. For each actual provider in the Texas ILINet, I estimate a reporting profile specified by 1) transition probabilities between reporting and non-reporting (Markov) states, and 2) the constants, modeling noise in the weekly ILI reports. To generate a mock provider in a specified zip code, I select a uniformly random reporting profile out of all reporting profiles estimated from existing providers. The generated mock providers are thereby given reporting characteristics typical of existing providers. I can then generate an ILI report time series for a mock provider, by 1) generating reports only during reporting weeks, and 2) calculating reports using equation (1) with the constants given in the provider's reporting profile and estimates of Percent-ILI(*i*) for the mock provider's zip code.

I select providers from pools consisting of a single mock provider from each zip code. Zip codes offer a convenient spatial resolution, because they have geographic specificity and are recorded in both the Texas ILINet and hospital discharge data. The optimization algorithm is not aware of a mock provider's reporting profile when the provider is selected (discussed below).

Provider Selection Optimization

The final step in the ILINet design method is selecting an optimized subset of providers from the mock provider pool. I seek the subset that most effectively predicts a target time series (henceforth, goal), as measured by the coefficient of determination (R^2) from a least squares multilinear regression to the goal from the report time series for all providers in the subset. Specifically, the objective function is given by

$$R^2(G,S) = \frac{Var(G) - Var\left(G - \sum_{i \in S} \alpha_i P_i\right)}{Var(G)},$$

Where G is the goal random variable; S is a subset of the mock provider pool; P_i are provider reports for provider i; and the α_i are the best multilinear regression coefficients (values that minimize the second term in the numerator).

There are several advantages to this objective function. First, it allows one to optimize an ILINet for predicting a particular random variable. Here, I set the goal to be state-wide influenza-related hospitalizations for Texas. This method can be applied similarly to design surveillance networks that predict, for example, morbidity and/or mortality within specific age groups or high risk groups.

Second, the objective function is submodular in the set of providers, S (Das and Kempe 2008), implying generally that adding a new provider to a small network will improve performance more than adding the provider to a larger network. The submodular property enables computationally efficient searches for near optimal networks and guarantees a good level of performance from the resulting network (Nemhauser et al. 1978). Without a submodular objective function, optimization of a k provider ILINet may require an exhaustive search of all subsets of k providers from the provider pool, which

quickly becomes intractable. For example, an exhaustive search for the optimal 200 provider Texas ILINet from the pool of approximately 2000 mock providers would require roughly 10^{660} regressions.

Taking advantage of the submodular property, I rapidly build high performing networks (with *k* providers) according to the following algorithm:

1. Let *P* be the entire provider pool, *S* be the providers selected thus far, and f(*S*) be a submodular function in *S*. I begin without any providers in *S*.

2. Repeat until there are *k* providers in *S*:

This is guaranteed to produce a network that performs within a fraction of $1-\frac{1}{e}$ of the optimal network (Das and Kempe 2008). The submodularity property also allows one to compute a posterior bound on the distance from optimality, which is often much better than $1-\frac{1}{e}$. Finally, even if implemented naively, the algorithm only requires approximately $10^{5.6}$ regressions to select 200 providers from a pool of 2000.

When optimizing, it is important to consider potential noise (underreporting and discrepancies between provider reports and actual ILI activity in the zip code). However, I assume that one cannot predict the performance of a particular provider before the provider is recruited into the network. To address this issue, the optimization's objective function is an expectation over the possible provider re- porting profiles. Specifically, I define ξ as a random variable describing the provider reporting profile for the entire pool of mock providers. If $\hat{\xi}$ is a specific reporting profile, then the R^2 objective function can be written as

$$R^2(G,S,\hat{\xi}) = \frac{Var(G) - Var\left(G - \sum_{i \in S} \alpha_i P_i(\hat{\xi})\right)}{Var(G)}.$$

To design the ILINet, I solve the following optimization problem

$$\max_{S \subseteq P} E_{\bar{\xi}}[R^2(G,S,\xi)].$$

The objective function is a convex combination of submodular functions, and thus is also submodular. This allows me to use the above algorithm along with its theoretical guarantees to design ILINets using a realistic model of reporting practices and informational stochasticity, without assuming that the designer knows the quality of specific providers *a priori*.

Maximal Coverage Model

I implemented the Maximal coverage model (MCM) following Polgreen et al. (2009). Briefly, a greedy algorithm was used to minimize the number of people in Texas who live outside a pre-defined coverage distance, C, of at least one provider in the selected set, S. A general version of this algorithm was developed by Church and Re Velle (1974) to solve this class of MCM's, As per Polgreen et al. (2009), I assumed that the population density of each zip code exists entirely at the geographic center of the zip code and used Euclidean distance to measure the distance between zip codes. Using a matrix of inter-zip code distances I select providers iteratively, choosing zip codes that cover the greatest amount of population density within the pre-defined coverage distance, C. I considered C = 5, 10, 20, and 25 miles, and found that C = 20 miles yielded the most informative networks.

Naive Methods

I used two naive methods to model common design practices for state-level provider-based surveillance networks.

1. *Greedy selection by population density* - All zip codes were ordered by population density and added to the provider pool P. Providers are then moved from P to the selected set S from highest to lowest density. The algorithm stops when S reaches a pre-determined size or P is empty.

2. *Uniform random by population size* - Zip codes are randomly selected from P and moved to S with a probability proportional to their population size. The algorithm proceeds until either S reaches a pre-determined size or P is empty.

Principal Component Analysis of Hospitalizations

To analyze similarities in ILI hospitalizations across different zip codes, I apply principal component analysis (PCA) (Jolliffe 2002). Specifically, I perform PCA on the centered (mean zero), standardized (unit variance) hospitalization time series of all zip codes in Texas. I first compute a time series for the first principal component, and then compute an R^2 for each zip code, based on a linear regression from the first principal component to the zip code's centered, standardized hospitalizations. Zip codes with high R^2 values have hospitalization patterns that exhibit high temporal synchronicity with the first principal component.

Out-of-sample Validation

To validate the method, I first use submodular optimization to create a provider network of 200 providers, using only data from 2001 to 2007, and then evaluate the

performance of the network in predicting 2008 influenza-like hospitalizations. Specifically, after creating the 200-provider network I use actual hospitalization data and mock provider reports for the 2001-2007 period to fit a multilinear regression model of the form $G^{train}(t) = \sum_{i \in S^{train}} \alpha_i^{train} P_i^{train}(t-2)$ where G^{train} is the time series of state-wide influenza-like hospitalizations at week t for weeks in 2001 to 2007, $P_i^{train}(t-2)$ is the mock report time series of provider i during week $t - 2$ for weeks in 2001 to 2007, and α_i^{train} is the best multilinear regression coefficient associated with provider i.

I then use the estimated multilinear regression function to forecast state-wide influenza-like hospitalization during 2008 from mock provider reports of 2008, and compare these forecasts to actual 2008 hospitalization data. This simulates a real-world prediction, where only historical data is available to create the provider network (S^{train}) and estimate the prediction function (α_i^{train}), and then the most recent provider reports (P_i^{2008}) are used to make predictions. I evaluate the 2008 predictions using a variance reduction measure similar to R^2, except that the multilinear prediction model uses coefficients estimated from prior data, as given by

$$\hat{R}^2(G^{2008}, S^{train}) = \frac{Var(G^{2008}) - Var\left(G^{2008} - \sum_{i \in S^{train}} \alpha_i^{train} E_\xi[P_i^{2008}(\xi)]\right)}{Var(G^{2008})},$$

where G^{2008} is the hospitalization time series in 2008, ξ is the provider noise profile, and $P_i^{2008}(\xi)$ are the mock provider reports in 2008. Importantly, I first calculate an expected value for the provider reports $P_i^{2008}(\xi)$, given the noise profiles ξ, before calculating $\hat{R}^2$.

References

Ahuja M, Anders F. 1976. A genetic concept of the origin of cancer, based in part upon studies of neoplasms in fishes. Prog Exp Tumor Res 20: 380 - 397.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control 119: 716 - 723.

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Current Opinion in Plant Biology 15: 140-146.

Barringer BC. 2007. Polyploidy and self-fertilization in flowering plants. American Journal of Botany 94: 1527-1533

Bateson W. 1909. Heredity and variation in modern lights. In: Seward A, editor. Darwin and Modern Science: Cambridge University Press. p. 85-101.

Beaumont M, Zhang W, Balding D. 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162: 2025-2035.

Bollback J. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7: 88.

Brownstein JS, Freifeld CC, Reis BY, Mandl KD. 2008. Surveillance Sans Frontires: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med 5: e151.

Bull JJ, Charnov EL. 1985. On Irreversible Evolution. Evolution 39: 1149-1155.

Bush, GW. 2007. Homeland security presidential directive 21: public health and medical preparedness. Available: http://www.dhs.gov/xabout/laws/gc_1219263961449.shtm. Accessed February 13th 2012.

Butler A, Trono D, Beard R, Fraijo R, Nairn R. 2007. Melanoma susceptibility and cell cycle genes in Xiphophorus hybrids. Mol. Carcinog. 46: 685-691.

Carrat F, Flahault A, Boussard E, Farran N, Dangoumau L, Valleron AJ. 1998. Surveillance of influenza- like illness in france. The example of the 1995/1996 epidemic. J Epidemiol Community Health 52: 32S-38S.

Cattani MV, Presgraves DC. 2009. Genetics and lineage-specific evolution of a lethal hybrid incompatibility between Drosophila mauritiana and its sibling species. Genetics 181: 1545-1555.

Censusgov. 2002. Census 2000 US Gazetteer Files. http://www2.census.gov/census_2000/ datasets/ Accessed February 13th 2012.

Church R, ReVelle C. 1974. The maximal covering location problem. Papers in Regional Science 32: 101-118.

Clothier H, Turner J, Hampson A, Kelly H. 2006. Geographic representativeness for sentinel influenza surveillance: implications for routine surveillance and pandemic preparedness. Aust NZ J Public Health 30: 337-341.

Cowling BJ, Wong IOL, Ho LM, Riley S, Leung GM. 2006. Methods for monitoring influenza surveillance data. Int J Epidemiol 35: 1314-1321.

Coyne J. 1992. Genetics and speciation. Nature 355: 511-515.

Cui R, Schumer M, Kruesi K, Walter RB, Andolfatto P, Rosenthal GG. 2013. Data from: Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes: Dryad Data Repository.

Cui R, Schumer M, Kruesi K, Walter RB, Andolfatto P, Rosenthal GG. 2013. Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. Evolution 67: 2166-2179.

Cutter AD. 2012. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. Trends in ecology & evolution 27: 210-219.

Das A, Kempe D. 2008. Algorithms for subset selection in linear regression. In: Proceedings of the 40th annual ACM symposium on Theory of computing. New York: ACM. pp. 45–54.

Deckers JG, Paget WJ, Schellevis FG, Fleming DM. 2006. European primary care surveillance networks: their structure and operation. Family Practice 23: 151-158.

Dobzhansky T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in Drosophilia pseudoobscura hybrids. Genetics 21: 113-135.

Fawcett JA, Maere S, Van der Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proceedings of the National Academy of Sciences USA 106: 5737-5742.

Fernandez A. 2010. A cancer-causing gene is positively correlated with male aggression in Xiphophorus cortezi. Journal of evolutionary biology 23: 386-396.

Fernandez A, Bowser P. 2010. Selection for a dominant oncogene and large male size as a risk factor for melanoma in the Xiphophorus animal model. Molecular Ecology 19: 3114-3123.

Fernandez AA, Morris MR. 2008. Mate choice for more melanin as a mechanism to maintain a functional oncogene. Proceedings of the National Academy of Sciences 105: 13503-13507.

FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution 3: 1084-1092.

Fleming D, Zambon M, Bartelds A, de Jong J. 1999. The duration and magnitude of influenza epidemics: A study of surveillance data from sentinel general practices in england, wales and the netherlands. Eur J Epidemiol 15: 467-473.

Franck D, Dikomey M, Schartl M. 2001. Selection and the maintenance of a colour pattern polymorphism in the green swordtail (Xiphophorus helleri). Behaviour 138: 467-486.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4: 177-183.

Gao P, Heller N, Walker W, Chen C, Moller M, Plunkett B, Roberts M, Schleimer R, Hopkin J, Huang S. 2004. Variation in dinucleotide (GT) repeat sequence in the first exon of the STAT6 gene is associated with atopic asthma and differentially regulates the promoter activity in vitro. Journal of medical genetics 41: 535-539.

Gavrilets S. 2003. Perspective: Models of speciation: What have we learned in 40 years? Evolution 57: 2197-2215.

Gavrilets S. 2004. Fitness landscapes and the origin of species. Princetion NJ: Princetion University Press.

Goldblatt P. [ed.] 1981. Index to plant chromosome numbers 1975-1978. Monographs in Systematic Botany. Missouri Botanical Garden, St. Louis USA.

Googleorg 2003. Explore flu trends - United States. http://www.google.org/flutrends/us/#US-TX. Accessed February 13th 2012.

Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinki M, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014.

Gordon M. 1931. Hereditary basis of melanosis in hybrid fishes. American Journal of Cancer 15: 1495-1523.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: Evo Devo and the genetics of adaptation. Evolution 61: 995-1016.

Huelsenbeck JP, Rannala B, Masly J. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288: 2349 - 2350.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences USA 104: 19369-19374.

Jiang X, Wallstrom G, Cooper GF, Wagner MM. 2009. Bayesian prediction of an epidemic curve. J Biomed Inform 42: 90 - 99.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97-100.

Jolliffe I. 2002. Principal Component Analysis. New York: Springer series in statistics.

Kallman K, Kazianis S. 2006. The genus Xiphophorus in Mexico and Central America. Zebrafish 3: 271-285.

Kang J, Schartl M, Meyer A. 2013. Comprehensive phylogenetic analyses of the genus Xiphophorus: hybrid origin of a swordtail fish, Xiphophorus monticolus. BMC Evolutionary Biology 13: 25 - 44.

Kazianis S, Coletta LD, Morizot DC, Johnston DA, Osterndorff EA, Nairn RS. 2000. Overexpression of a fish CDKN2 gene in a hereditary melanoma model. Carcinogenesis 21: 599-605.

Kazianis S, Gutbrod H, Nairn RS, McEntire BB, Della Coletta L, Walter RB, Borowsky RL, Woodhead AD, Setlow RB, Schartl M, et al. 1998. Localization of a CDKN2 gene in linkage group V of Xiphophorus fishes defines it as a candidate for the DIFF tumor suppressor. Genes, chromosomes, and cancer 22: 210-220.

Kazianis S, Khanolkar VA, Nairn RS, Rains JD, Trono D, Garcia R, Williams EL, Walter RB. 2004. Structural organization, mapping, characterization and evolutionary relationships of CDKN2 gene family members in Xiphophorus fishes. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 138: 291-299.

Kazianis S, Morizot D, Della Coletta L, Johnston D, Woolcock B, Vielkind J, Nairn R. 1999. Comparative structure and characterization of a CDKN2 gene in a Xiphophorus fish melanoma model. Oncogene 18: 5088-5099.

Kazianis S, Morizot DC, McEntire BB, Nairn RS, Borowsky RL. 1996. Genetic mapping in Xiphophorus hybrid fish: assignment of 43 AP-PCR/RAPD and isozyme markers to multipoint linkage groups. Genome Research 6: 280-289.

Khan AS, Fleischauer A, Casani J, Groseclose SL. 2010. The Next Public Health Revolution: Public Health Information Fusion and Social Networks. Am J Public Health 100: 1237-1242.

Kosswig C. 1928. Uber Kreuzungen zwischen den Teleostiern Xiphophorus helleri und Platypoecilus maculatus. Z Indukt Abstammungs- Vererbungsl 47: 150-158.

Lajeunesse MJ. 2009. Meta-analysis and the comparative phylogenetic method. American Naturalist 174: 369-381.

Lajeunesse MJ. 2011. phyloMeta: a program for phylogenetic comparative analyses with meta-analysis. Bioinformatics 27: 2603-2604.

Lande R, Schemske DW. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. Evolution 39: 24-40.

Langmuir A, Housworth J. 1969. A critical evaluation of influenza surveillance. Bull World Health Organ 41: 393-398.

Levin DA. 1975. Minority cytotype exclusion in local plant populations. Taxon 24: 35-43.

Levin DA. 1983. Polyploidy and novelty in flowering plants. American Naturalist 122: 1-25.

Levin DA. 2002. The role of chromosomal change in plant evolution. Oxford Univ. Press, New York.

Mable BK. 2004. Polyploidy and self-compatibility: is there an association? New Phytologist 162: 803-811.

Mayrose I, Barker MS, and Otto SP. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. Systematic Biology 59:132-144.

Mayrose I, Otto SP. 2010. A Likelihood Method for Detecting Trait-Dependent Shifts in the Rate of Molecular Evolution. Molecular Biology and Evolution 28: 759-770.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. Science 333: 1257.

Meierjohann S, Schartl M. 2006. From Mendelian to molecular genetics: the Xiphophorus melanoma model. Trends in genetics 22: 654-661.

Merbs SL, Sidransky D. 1999. Analysis of p16 (CDKN2/MTS-1/INK4A) alterations in primary sporadic uveal melanoma. Investigative ophthalmology & visual science 40: 779-783.

Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB, Sidransky D. 1995. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nature Medicine 1: 686-692.

Meyer A, Morrissey J, Schartl M. 1994. Recurrent origin of a sexually selected trait in Xiphophorus fishes inferred from a molecular phylogeny. Nature 368: 539-542.

Meyer A, Salzburger W, Schartl M. 2006. Hybrid origin of a swordtail species (Teleostei: Xiphophorus clemenciae) driven by sexual selection. Molecular Ecology 15: 721-730.

Meyers LA, Levin DA. 2006. On the abundance of polyploids in flowering plants. Evolution 60: 1198-1206.

Missouri Botanical Garden. 2005. Tropicos, botanical information system at the Missouri Botanical Garden - www.tropicos.org.

Mnatsakanyan ZR, Burkom HS, Coberly JS, Lombardo JS. 2011. Bayesian Information Fusion Networks for Biosurveillance Applications. J Am Med Inform Assoc 16: 855-863.

Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J. 2000. Clinical Signs and Symptoms Predicting Influenza Infection. Arch Intern Med 160: 3242-3247.

Moore RJ. [ed.] 1973. Index to plant chromosome numbers 1967-1971. Monographs in Systematic Botany. Missouri Botanical Garden, St. Louis USA.

Muller H. 1942. Isolating mechanisms, evolution and temperature. Biol Symp 6: 71-125.

Nemhauser GL, Wolsey LA, Fisher ML. 1978. An analysis of approximations for maximizing submodular set functions—i. Math Program 14: 265-294.

Nei M, Nozawa M. 2011. Roles of Mutation and Selection in Speciation: From Hugo de Vries to the Modern Genomic Era. Genome Biology and Evolution 3: 812-829.

Ordobas M, Zorrilla B, Arias P. 1995. Influenza in madrid, spain, 1991-92: validity of the sentinel network. J Epidemiol Community Health 49: 14-16.

Otto SP. 2007. The evolutionary consequences of polyploidy. Cell 131: 452-462.

Otto SP, and Whitton J. 2000. Polyploid incidence and evolution. Annual Review of Genetics 34: 401-437.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society (B) 255: 37-45.

Polgreen P, Chen Z, Segre A, Harris M, Pentella M, Rushton G. 2009. Optimizing influenza sentinel surveillance at the state level. Am J Epidemiol 170.

Pond SLK, Frost SDW. 2004. A Genetic Algorithm Approach to Detecting Lineage-Specific Variation in Selection Pressure. Molecular Biology and Evolution 22: 478-485.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. Nature Reviews Genetics 11: 175-180.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between. Nature 423: 715-719.

Quenel P, Dab W. 1998. Influenza a and b epidemic criteria based on time-series analysis of health services surveillance data. Eur J Epidemiol 14: 275-285.

Rath T, Carreras M, Sebastiani P. 2003. Automated detection of influenza epidemics with hidden markov models. In: Berthold M, Lenz H, Bradley E, Kruse R, Borgelt C, editors. Advances in Intelligent Data Analysis V, Berlin-Heidelberg: Springer. pp. 521-532.

Regneri J, Schartl M. 2011. Expression regulation triggers oncogenicity of xmrk alleles in the Xiphophorus melanoma system. Comparative Biochemistry and Physiology, Part C: 1-10.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology 61: 539-542.

Rosenthal GG, Garcia-De-Leon FJ. 2011. Speciation and hybridization. In: Schlupp I, Pilastro A, Evans J, editors. Ecology and Evolution of Poeciliid Fishes. Chicago, IL: University of Chicago Press. p. 109-119.

Schartl M. 1990. Homology of melanoma-inducing loci in the genus Xiphophorus. Genetics 126: 1083-1091.

Schartl M. 2008. Evolution of Xmrk: an oncogene, but also a speciation gene? BioEssays 30: 822-832.

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff J, Lesch K, et al. 2013. The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nature Genetics 45: 567-572.

Schluter D, Conte G. 2009. Genetics and ecological speciation. Proceedings of the National Academy of Sciences 106: 9955-9962.

Schumer M, Cui R, Boussau B, Walter RB, Rosenthal GG, Andolfatto P. 2012. An evaluation of the hybrid speciation hypothesis for Xiphophorus clemenciae based on whole genome sequences. Evolution 67: 1155-1168.

Soltis DE, Burleigh JG. 2009. Surviving the K-T mass extinction: New perspectives of polyploidization in angiosperms. Proceedings of the National Academy of Sciences USA 106: 5455-5456.

Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. Trends in Ecology & Evolution 14: 348-352.

Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant Speciation. New Phytologist 161: 173-191.

Soltis DE, Albert V, Leebens-Mack J, Bell C, Paterson A, Zheng C, Sankoff D, Depamphilis C, Wall P, Soltis PS. 2009. Polyploidy and angiosperm diversification. American Journal of Botany 96: 336-348.

Stebbins GL. 1950. Variation and evolution in plants. Columbia University Press, New York City, NY USA.

Stebbins GL. 1971. Chromosomal evolution in higher plants. Addison-Wesley, Reading, Masachusetts USA.

Stebbins GL. 1980. Polyploidy: future prospects. in W.H. Lewis [ed.], Polyploidy - Biological Relevance. 495-520. Springer, New York.

Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. Annals of the Missouri Botanical Gardens 72: 824-832.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. Theory Methods A7: 13 - 26.

Weis S, Schartl M. 1998. The macromelanophore locus and the melanoma oncogene Xmrk are separate genetic entities in the genome of Xiphophorus. Genetics 149: 1909-1920.

te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubešová M, and Pyšek P. 2012. The more the better? The role of polyploidy in facilitating plant invasions. Annals of Botany 109: 19-45.

Thompson WW, Shay DK, Weintraub E, Brammer L, Bridges CB, Cox NJ, Fukuda K. 2004. Influenza-associated hospitalizations in the united states. JAMA 292: 1333-1340.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumf MPH. 2008. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface: 1-16.

US Centers for Disease Control and Prevention. 2010. National biosurveillance strategy for human health. Available: http://www.cdc.gov/osels/ph_surveillance/bc.html. Accessed February 13th 2012.

Valdivia A, Lopez-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. 2010. Monitoring influenza activity in europe with google flu trends: comparison with the findings of sentinel physician networks – results for 2009-10. Euro Surveill 15: 1-6.

Vamosi JC, Dickinson TA. 2006. Polyploidy and diversification: in phylogenetic investigation in Rosaceae. International Journal of Plant Sciences 167: 349-358.

Viboud C, Boelle P, Carrat F, Valleron A, Flahault A. 2003. Prediction of the spread of influenza epidemics by the method of analogues. Am J Epidemiol 158: 996-1006.

Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312: 447-451.

Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. 2009. Interpreting Google flu trends data for pandemic H1N1 influenza: the New Zealand experience. Euro Surveill 14: 1-3.

Wittbrodt J, Adam D, Malitschek B, Mäueler W, Raulf F, Telling A, Robertson S, Schartl M. 1989. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing Tu locus in Xiphophorus. Nature 341: 415-421.

Wood, TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences USA 106: 13875-13879.

Wu CI, Ting CT. 2004. Genes and speciation. Nature Reviews Genetics 5: 114-122.

Yang P, Duan W, Lv M, Shi W, Peng X, Wang X, Lu Y, Liang H, Seale H, Pang X et al. 2009. Review of an influenza surveillance system, beijing, people's republic of china. Emerg Infect Dis 15: 1603-1608.